Javier Carrillo-Mondéjar
Manuel Gil Pérez
Sergio Pastrana
Ricardo J. Rodríguez
(Eds.)



X Jornadas Nacionales de Investigación en Ciberseguridad

Zaragoza, del 4 a 6 de junio de 2025

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD

Universidad Zaragoza
1474

uc3m | Universidad Carlos III de Madrid

Ricardo J. Rodriguez · Sergio Pastrana · Manuel Gil · Javier Carrillo (Eds.)

Actas de las X Jornadas Nacionales de Investigación en Ciberseguridad

X Jornadas Nacionales de Investigación en Ciberseguridad, JNIC 2025.
Zaragoza, del 4 a 6 de junio de 2025





Servicio de Publicaciones
Universidad Zaragoza

Editores

Ricardo J. Rodriguez
Universidad de Zaragoza

Sergio Pastrana
Universidad Carlos III de Madrid

Manuel Gil Pérez
Universidad de Murcia

Javier Carrillo
Universidad de Zaragoza

ISBN: 978-84-10169-61-6
Primera Edición, junio 2025
© Universidad de Zaragoza
Nota: Este documento toma como referencia las actas de JNIC 2024, disponibles aquí.

Esta obra se encuentra bajo una licencia Creative Commons CC BY-NC 4.0. Está permitida cualquier forma de copia, reproducción y redistribución del material de esta obra en cualquier medio o formato, pero no se permite un uso comercial de la obra original ni la generación de obras derivadas. Puede acceder al texto completo de la licencia en el siguiente enlace: https://creativecommons.org/licenses/by-nc/4.0/deed.es

Patrocinadores

Patrocinadores Oro







Patrocinadores Plata







Comité técnico

Comité del programa científico

Presidentes

Javier Carrillo Mondéjar	Universidad de Zaragoza
Manuel Gil Pérez	Universidad de Murcia

Comité

Jorge Blasco Alís	Universidad Politécnica de Madrid
Pino Caballero-Gil	Universidad de La Laguna
Andrés Caro	Universidad de Extremadura
Javier Carrillo Mondéjar	Universidad de Zaragoza
Juan Manuel Castelo Gómez	Universidad Politécnica de Madrid
Jesús Díaz-Verdejo	Universidad de Granada
Santiago Escobar Román	Universidad Politécnica de Valencia
Josep Lluís Ferrer Gomila	Universitat de les Illes Balears
Jose María de Fuentes	Universidad Carlos III de Madrid
Joaquin Garcia-Alfaro	Institut Polytechnique de Paris
Luis Javier García Villalba	Universidad Complutense de Madrid
Manuel Gil Pérez	Universidad de Murcia
Lorena González Manzano	Universidad Carlos III de Madrid
Luis Hernández Encinas	Institute of Physical and Information Technologies Spanish National Research Council (CSIC)
Javier López	Universidad de Málaga
Gregorio López López	Universidad Pontificia Comillas
Agustín Martín Muñoz	Institute of Physical and Information Technologies Spanish National Research Council (CSIC)
José Luis Martínez Martínez	Universidad de Castilla-La Mancha
Gregorio Martínez Pérez	Universidad de Murcia
David Megias	Universitat Oberta de Catalunya
Pantaleone Nespoli	Universidad de Murcia
Raúl Orduña	VICOMTECH
Inés Ortega Fernández	GRADIANT
Luis Panizo	Universidad de León
Alojas Pasic	Eviden
Razvan Raducu	Universidad de Zaragoza
Ricardo J. Rodríguez	Universidad de Zaragoza
José Roldán	Universidad de Zaragoza
M. Carmen Romero-Ternero	Universidad de Sevilla
David G. Rosado	Universidad de Castilla-La Mancha
Sergio Ruiz Villafranca	Universidad de Castilla-La Mancha

Luis Enrique Sánchez Crespo	Universidad de Castilla-La Mancha
Daniel Uroz	Universidad de Zaragoza
Ángel Jesús Varela Vaca	Universidad de Sevilla
José M. Vázquez-Naya	Universidade da Coruña
Manuel Veiga	Universidad de Vigo
Urko Zurutuza Ortega	Mondragon Unibertsitatea
Victor Villagrá	Universidad Politécnica de Madrid

Comité del programa de transferencia

Presidentes

Cristina Alcaraz	Universidad de Málaga
Juan González Martínez	GRADIANT

Comité

Saioa Arrizabalaga	CEIT
Ana Isabel Ayerbe	TECNALIA
Francisco Luis Benítez Martínez	FIDESOL
Pino Caballero-Gil	Universidad de La Laguna
Juan Caubet	EURECAT
Rodrigo Diaz	i2CAT
Juan Díez	INCIBE
Eduardo Fernández-Medina	Universidad de Castilla La Mancha
David González	IKERLAN
Ricardo J. Rodríguez	Universidad de Zaragoza
Urko Zurutuza Ortega	Mondragon Unibertsitatea

Comité del programa de formación

Presidentes

Ana Isabel González-Tablas	Universidad Carlos III de Madrid
María Teresa Gómez López	Universidad de Sevilla

Comité

Isaac Agudo Ruiz	Universidad de Málaga
Mar Ávila Vegas	Universidad de Extremadura
María Isabel Barceló Villanueva	Universitat de les Illes Balears
Yolanda Blanco-Fernández	Universidad de Vigo
Noemí de Castro García	Universidad de León
David G. Rosado	Universidad de Castilla La Mancha
Iñaki Garitano Garitano	Mondragon Unibertsitatea
Xavier Larriva-Novo	Universidad Politécnica de Madrid
María de los Llanos Tobarra Abad	UNED
Roberto Magán-Carrión	Universidad de Granada
Victoria Alicia Marcos Sbarbaro	INCIBE
Óscar Mogollón Gutiérrez	Universidad de Extremadura
María Magdalena Payeras Capellà	Universitat de les Illes Balears
Antonio Robles Gómez	UNED
José Carlos Sancho Núñez	Universidad de Extremadura
Adriana Suárez Corona	Universidad de León
Ángel Jesús Varela Vaca	Universidad de Sevilla

Comité de organización

Comité de organización

Presidentes

Ricardo J. Rodríguez	Universidad de Zaragoza
Sergio Pastrana	Universidad Carlos III de Madrid

Miembros

Álvaro Alesanco	Universidad de Zaragoza
Pedro Álvarez	Universidad de Zaragoza
María Carmen Cámara Nuñez	Universidad Carlos III de Madrid
Alejandro de la Cruz	Universidad Carlos III de Madrid
José García Moros	Universidad de Zaragoza
Virginia Giménez Mazas	Universidad de Zaragoza
Ana Isabel González-Tablas	Universidad Carlos III de Madrid
María Porfiria	Universidad Carlos III de Madrid

Índice general

Conferencias invitadas 3

Oracle Parfait: Real-World Vulnerability Detection at Scale and AI-Assisted Patch Generation 3

¿Cómo atacan los que atacan en el 2025? 5

Programa científico 9

Sesión Investigación I: Privacidad, regulación y transparencia 9

E. García-Arias, A. Robles-Gómez, R. Pastor-Vargas, L. Tobarra, P. Vidal-Balboa and M. Méndez-Suárez
Dataset de Licencia y Términos de Uso de Apps Móviles y Análisis de Riesgos de Privacidad 9

P. A. Criado-Lozano, M. M. Martínez-González, A. Pérez-Fuente and A. Aparicio
El espejismo de la transparencia: Discrepancias en las declaraciones de privacidad en apps móviles . . . 18

A. Pérez-Fuente, M. M. Martínez-González, A. Aparicio and P. A. Criado-Lozano
¿Es Health Connect la mejor garantía de Seguridad y Privacidad para los datos de salud? 26

F. J. Rodríguez Lera, Y. Pita Lorenzo, D. Sobrín Hidalgo, L. Fernández Becerra, I. González Fernández and J. M. Guerrero Hernández
Aportes para el cumplimiento del Reglamento (UE) 2024/1689 en robótica y sistemas autónomos . . . 34

M. Buitrago-Perez, J. M. de Fuentes, O. Tejerina, L. Gonzalez-Manzano, N. Anciaux and L. Ibanez-Lissen
XAI for safe AI: technical-legal misconceptions and misalignments . . . 42

D. Sobrín-Hidalgo, Á. M. Guerrero-Higueras, F. J. Rodríguez-Lera and V. Matellán-Olivera
Definiendo el concepto de explicación: propuesta de una taxonomía para explicabilidad en robótica . . . 50

Sesión Transferencia I 59

Á. Barriga Rodríguez, Á. Hernández Martín, B. González Sánchez, D. Santos Anés, J. Nieto Acero and S. García Gil
Protección de Información Biométrica con Fuzzy Vault y Cifrado Homomórfico . . . 59

A. Fernández Isabel, I. Martín De Diego, E. López Cano, R. Rodríguez Fernández, J. Garcia-Ochoa, R. R. Ravines, O. López Espinosa and J. Puigbó Sanvisens
Dynamic Industrial Cyber risk Modelling based on Evidence (DICYME) . . . 64

C. Fernández Rosales, C. M. Alba Jiménez, R. Morillas, R. Román Aguilar, L. Vigoya and A. González Muñiz
CICERO-DARA Prototipo de informes de seguridad para entidades . . . 68

Sesión Investigación II: IA en ciberseguridad 73

D. Escudero García, Ó. Mogollón-Gutiérrez, J. C. Sancho Núñez and N. DeCastro-García
Optimización de la detección de intrusiones mediante estrategias adaptativas con aprendizaje por refuerzo . . . 73

I. Gamiz, C. Regueiro and E. Jacob
Hacia un Aprendizaje Federado Confiable: Análisis y extensión de PRoT-FL 82
J. Fúster, R. Gonzalez, G. Siracusano, D. Sanvito and R. Bifulco
Design of a GenAI-Powered Platform for Cyber Threat Intelligence Analysis 90
F. J. Blanco-Jiménez, F. González-López, A. Casares-Andrés and P. García-Teodoro
Clasificación Automática de Canales Telegram con BERTopic y DeepSeek 98
M. Marín-Caballero, M. Betancourt Alonso, D. Díaz-López, Á. L. Perales Gómez, P. Nespoli and G. Martínez Pérez
Anticipating Adversary Behavior in DevSecOps Scenarios through Large Language Models 105
M. Moreno, X. Sáez-de-Cámara, A. Urbieta and M. Iturbe
Leveraging LLMs for Automated IDS Rule Generation: A Novel Methodology for Securing Industrial Environments 113

Sesión Transferencia II 121
J. A. Ruipérez Valiente, F. Gómez Mármol and F. J. García Clemente
SCORPION Cybertech: spin-off dedicada a la formación en ciberseguridad con un Cyber Range 121
B. Pérez Vázquez, A. Fontán Santamaría and J. Fuentes Rodríguez
Detección de desinformación en redes sociales: aplicación y validación en un entorno real . 126
A. Buitrago López, T. Garcia De Alcaraz Ruiz, M. Fernández Tárraga, A. D. Cayuela Tudela, F. Sánchez González, J. Pastor-Galindo, J. A. Ruipérez-Valiente and G. Martínez Pérez
EUCINF: EUropean Cyber and INFormation warfare toolbox 130
J. Román Martínez, D. Triana Robles, M. El Oualidi Charchmi, I. Salamanca Estévez, N. DeCastro-García
Generación de títulos en español para la alerta temprana de vulnerabilidades mediante Inteligencia Artificial Generativa 134

Sesión Investigación III: Ciberseguridad en IoT y entornos inteligentes 141
L. M. García-Sáez, S. Ruiz-Villafranca, J. Roldán-Gómez, J. Carrillo-Mondéjar and J. L. Martínez
Estrategia de Agregación Ponderada para Sistemas IDS Federados en IoT 141
M. Rashed, C. García-Rubio and C. Campo
Studying MQTT Malicious Traffic: Methods and Challenges 150
J. L. Ruiz Catalán, F. Martinez, Á. Suárez-Bárcena, A. Santos-Olmo, D. G. Rosado and L. E. Sánchez Crespo
Análisis y Gestión de riesgos de ciberseguridad en entornos IoT: El patrón MARISMA-IoT 157
J. Muñoz-Calle, J. Diaz-Verdejo, R. Estepa Alonso and A. Estepa Alonso
Análisis comparativo de las capacidades de SIDS en un escenario SmartHome real 165
L. Mallen, J. A. Sanguesa, J. Roldán-Gómez, V. Torres-Sanz and F. J. Martinez
SecureAutoLoRa: Proceso de Registro Automático Seguro para Dispositivos LoRaWAN . . 173
D. F. Bustamante V., L. E. Sánchez, D. G. Rosado, A. Santos-Olmo and E. Fernández-Medina
QUILLAQUA: Hacia un marco sostenible para entornos hidropónicos seguros y eficientes en la Agricultura 4.0 181

Sesión Investigación IV: Seguridad en sectores estratégicos 189
P. Tuset-Peiró, M. Pilgermann, J. Pegueroles and X. Vilajosana
Assessing cybersecurity of Internet-facing medical IT systems in Germany & Spain using OSINT tools 189
J. A. Guillén-Mass, R. A. Rodríguez-Gómez and R. Magán-Carrión
From Theory to Practice: Assessing the Suitability of Testbeds in Industrial Cybersecurity 198

M. Franco de la Peña, Á. L. Perales Gómez and L. Fernández Maimó
Un módulo de explicabilidad basado en los valores Shapley para frameworks de detección de anomalías sobre datos secuenciales: Temporal SHAP . . . 206
E. T. Martínez Beltrán, M. Fernández Llamas, A. Zine Boujemaoui, G. Bovet, G. Martínez Pérez and A. Huertas Celdrán
Aprendizaje Federado Descentralizado y Conciencia Situacional para el Reconocimiento Aéreo Militar Seguro y Resiliente . . . 214
A. Feijóo-Añazco, A. López Martínez, D. Díaz-López, Á. L. Perales Gómez, P. Nespoli and G. Martínez Pérez
Toward a Military Smart Cyber Situational Awareness (CSA) . . . 222
F. Martínez, J. L. Ruiz Catalán, L. E. Sánchez, A. Santos-Olmo, D. G. Rosado and E. Fernández-Medina
Ciberseguridad Marítima en la era digital: Evaluación y validación de la planificación en POSEIDON . . . 230

Sesión Investigación V: Criptografía clásica y post-cuántica 239
D. Perez Perez and S. Paguada
Exploration of Deep Learning Convolutional Models Applied to Side-Channel Analysis . . 239
A. Blesa Gascón, F. J. Serón Arbeloa and D. Sanz Fuertes
Mejora del algoritmo de encriptado DRPE para CGH de fase mediante barajado aleatorio de píxeles . . . 248
J. Garcia-Diaz, D. Escanez-Exposito, P. Caballero-Gil and J. Molina-Gil
QZKP Basada en Codificación Conjugada para Autenticación de Usuarios . . . 255
L. Syne, É. Pérez-Ramos, C. Hernández-Goya and P. Caballero-Gil
Towards IoT Security: A Hybrid Encryption Scheme based on Lightweight and Post-Quantum Cryptography . . . 262
M. Rodriguez-Vega and P. Caballero-Gil
Protección de Inferencias en Redes Neuronales Profundas usando Cifrado Homomórfico . . 269
J. Padilla Pío, C. Caballero-Gil, J. Molina-Gil and M. Yung
Diseño de un sistema de votación electrónica basado en Blockchain y cifrado homomórfico 276

Sesión Investigación VI: Ciberamenazas y detección avanzada 283
J. Román Martínez, D. Triana Robles, M. El Oualidi Charchmi, I. Salamanca Estévez and N. DeCastro-García
Generación de títulos en español para la alerta temprana de vulnerabilidades mediante Inteligencia Artificial Generativa . . . 283
C. Sánchez-Zas, S. García Sánchez, X. Larriva-Novo, S. Solera-Cotanilla, O. Jover-Walsh and V. A. Villagrá
ARGOS: Herramienta para la simulación de rutas de ataque en entornos de conciencia cibersituacional . . . 291
E. Moreno Sánchez, T. Marinaro, R. Williams, M. Patrignani, H. Nemati, R. Guanciale and M. Guarnieri
(Work in Progress) Automated Synthesis of Instruction-Centric Leakage Contracts . . . 299
J. M. Tejada-Triviño, E. Castillo-Fernández, P. García-Teodoro and J. A. Gómez-Hernández
Herramienta Automatizada para Identificar Datos Personales en Exfiltraciones por Ransomware . . . 307
Á. Sánchez-Corchado, L. Gutiérrez-Galeano and J. J. Domínguez-Jiménez
Sistema Generador de Conjuntos de Datos con Ciberataques . . . 315
P. Beltrán-Lopez, M. Gil Pérez and P. Nespoli
Honeypot Adaptativa y Dinámica Impulsada por la IA para Estrategias de Engaño . . . 323

Sesión Investigación VII: Plataformas y marcos de defensa cibernética 331

V. Figueroa, L. E. Sánchez Crespo, A. Santos-Olmo, D. G. Rosado and E. Fernández-Medina
Hacia una ciberseguridad holística en e-Gobierno: Extensión de X-Road y MARISMA en el framework GAUCHO . . . 331
D. Alvarez, L. Perez, A. Mateo, X. Larriva-Novo, M. Alvarez-Campana and V. A. Villagrá
Diseño y generación de un conjunto de datos para entrenamiento de modelos de prevención y detección de amenazas internas: SPEDIA dataset . . . 340
E. Genua, M. Iturbe, I. Garitano, X. Etxezarreta, A. Aguirre and U. Zurutuza
Port2Vec: Generación de embeddings a partir de descripciones de servicios de red para su uso en Inteligencia Artificial aplicada a la Ciberseguridad . . . 348
A. M. López Gómez, A. A. Rodríguez Rodríguez, O. Tejerina Rodríguez, J. Inza Aldaz and G. López López
Verificación de edad en línea basada en extensión de navegador y pruebas de conocimiento cero . . . 356
A. G. Márquez Trujillo, Á. J. Varela-Vaca, M. T. Gómez López, A. M. Reina Quintero and R. Corchuelo
Mejorando la seguridad de la cadena de suministro de software . . . 364
A. Sánchez-Matas, P. Escribano Ruíz, D. Díaz-López, Á. L. Perales Gómez, P. Nespoli and G. Martínez Pérez
Simulating Cyberattacks through a Breach Attack Simulation (BAS) Platform empowered by Security Chaos Engineering (SCE) . . . 372

Sesión Formación I 381
A. Bofill González, D. Gordillo Burrero, A. López Trigo and A. Caro Lindo
Laboratorio de aprendizaje de Inyecciones SQL en entornos de Hacking Ético . . . 381
A. J. Di Bartolo, A. López Trigo, D. Gordillo Burrero, A. Bofill González and J. C. Sancho Núñez
De novatos a expertos: Creando una comunidad estudiantil para fomentar la ciberseguridad 390
B. Abbaoui Toufiq, M. Homaei, M. J. Sequera Fernández and M. M. Ávila Vegas
Estrategias de securización en el Cloud . . . 398
D. Cabuya-Padilla, D. Díaz-López and C. Castaneda-Marroquín
Hybrid Tabletop Exercise (TTX) based on a Mathematical Simulation-based Model for the Maritime Sector . . . 406
M. Salas Heras and A. I. González-Tablas
The QoQo Docker: A Docker Environment for Learning Quantum and Post-Quantum Cryptography in Python . . . 414

Sesión Investigación VIII: Concienciación y temas emergentes en ciberseguridad 421
M. P. Marco Francia, J. C. Ortiz-Pradillo, M. T. Montero Ruiz and A. L. Sandoval Orozco
Ciberseguridad y delitos sexuales en víctimas menores de dieciséis años . . . 421
G. Pinedo Terrados and A. I. González-Tablas
Exploring the impact of Spanish-specific trained PassGPT model for password guessing . . 428
G. Gomez, K. van Liebergen, D. Sanvito, G. Siracusano, R. Gonzalez and J. Caballero
(Work-in-progress) Measuring Data Pollution in Cryptocurrency Abuse Reporting Services 436
M. Calonge Fabregat, J. A. Medina Merodio and J. Hernández Castro
Una propuesta para mejorar la seguridad de las hardware wallets . . . 442
D. Povedano Álvarez, A. L. Sandoval Orozco and L. J. García Villalba
Técnica de Fusion Audio-Visual para la Detección de Contenido Sensible en Videos . . . 449
V. M. López Madejska, S. López Bernal, G. Martínez Pérez and A. Huertas Celdrán
Recreación de ceguera inducida por un ciberataque neuronal en una topología neuronal realista . . . 457

Sesión Formación II 465

L. Tobarra, A. Robles-Gómez, R. Pastor Vargas and R. Hernández
Contenido Interactivo y Accesible para el Aprendizaje de Ciberseguridad 465
J. Muñoz-Calle, F. J. Fernández-Jiménez, R. Estepa, V. Mayor and J. M. Garcia-Campos
Experiencias de formación en ciberseguridad usando portales cautivos 473
S. Frías-Guadalupe, P. Paredes and C. Castaneda-Marroquín
Cibersapiens: Gamifying Cybersecurity Education to Engage Secondary Students 481
É. Pérez-Ramos, C. Hernández Goya and P. Caballero-Gil
Gamificación con Genially para aprender criptografía post-cuántica 490

Programa científico: pósters 501

Sesión Pósters I: Detección de amenazas y ciberataques emergentes 501
J. Carrillo-Mondéjar, G. Suarez-Tangil, A. Costin and R. J. Rodríguez
A Review of "Exploring Shifting Patterns in Recent IoT Malware" 501
A. Barreiro-Morante, D. García-Algora, A. C. del Rio-Álvarez and J. Blasco
TRAKR: A framework to collect, process and analyze diverse APT files and reports 504
A. de la Cruz and S. Pastrana
A review of: Understanding Crypter-as-a-Service in a popular underground marketplace . 509
A. Rodriguez Barredo-Valenzuela, S. Pastrana Portillo, N. Vallina-Rodriguez and G. Suarez-Tangil
A Review Of: Reversing the Virtual Maze - An Overview of the Technical and Methodological Challenges for Metaverse App Analysis 511
J. García-Ochoa, A. Fernández-Isabel, I. Martín de Diego, C. Contreras, R. R. Ravines and O. López
Defining the Basal Attractiveness Concept for Cybercriminals 513
M. Homaei, V. González Morales, Ó. Mogollón-Gutiérrez and A. Caro
The Dark Side of Digital Twins: Adversarial Attacks on AI-Driven Water Forecasting . . . 518
D. Galán, V. García, S. Escobar, C. Meadows and J. Meseguer
A Review of: Protocol Dialects as Formal Patterns 523
J. Á. Fernández Carrasco, I. Amonarriz Pagola, R. Orduna Urrutia and R. Román
A review of: CYBERSHIELD: A Competitive Simulation Environment for Training AI in Cybersecurity 525
E. Iturbe, O. Llorente-Vazquez, A. Rego, E. Rios and N. Toledo
Una revisión de: "Desatando la Inteligencia Artificial ofensiva: Generación automática de código de técnicas de ataque" 527
C. M. Braga, M. A. Serrano and E. Fernández-Medina
Hacia la integración autónoma de ATT&CK y marcos de seguridad mediante el uso de Agentic AI 529

Sesión Pósters II: Privacidad, datos sensibles y tecnologías cuánticas 535
X. Martínez-Luaña, M. Fernández-Veiga and R. P. Díaz-Redondo
A review of "Privacy-aware Berrut Approximated Coded Computing applied to Federated Learning" 535
F. Kabir, T. Koohpayeh Araghi and D. Megías
A review of "Privacy-preserving protocol for high-frequency smart meters using reversible watermarking and paillier encryption" 538
J. Loureiro-Acuña, X. Martínez-Luaña, H. Padín-Torrente, G. Jiménez-Balsa, C. García-Pagán and I. Ortega-Fernandez
A review of: Enhancing Privacy in Federated Learning: A Practical Assessment of Combined PETs in a Cross-Silo Setting 540

D. Soler, I. Cillero, C. Dafonte, F. Nóvoa, M. Fernández-Veiga and A. Fernández Vilas
A Review of "QKDNetSim+: Improvement of the Quantum Network Simulator for NS-3" . 542
R. Díaz Bonete, M. I. García Cid and D. Domingo Martín
Deploying a Hybrid Quantum-Safe VPN for Secure Banking Communications 544
D. Blanco-Aza, A. Robles-Gómez, R. Pastor-Vargas, L. Tobarra, P. Vidal-Balboa and M. Méndez-Suárez
Análisis de Privacidad en Aplicaciones Móviles y Redes Sociales mediante Técnicas de IA 549
T. Koohpayeh Araghi and D. Megías
A Review of "Analysis and effectiveness of deeper levels of SVD on performance of hybrid DWT and SVD watermarking" 551
D. Escanez-Exposito, J. Correa-Marichal and P. Caballero-Gil
Una revisión de "Uso del aprendizaje basado en juegos y la computación cuántica para mejorar las competencias STEAM en la educación K-16" 553
L. M. Muñoz Pérez and R. Igual Catalán
Securización frente a decompilaciones: enfoque en aplicaciones móviles Android de Machine Learning 555
S. Escuder Folch, A. Calvo, N. Ortiz, J. Escrig and M. Compastié
A Review of of RBD24 : A Labelled Dataset with Risk Activities Using Log Application Data 560

Sesión Pósters III: Aplicaciones de IA en ciberseguridad 563
F. Jáñez-Martino, E. Fidalgo, R. Alaiz-Rodríguez, A. Martínez-Mendoza and E. Díez-del-Río
A review of comparative analysis of natural language processing models for spam email identification 563
A. Martínez-Mendoza, E. Fidalgo, E. Alegre, W. Tanveer and E. Díez-del-Río
A review of Enhancing CSIRT effectiveness with transformer-based cyber incident classification 566
A. Martínez-Mendoza, F. Jáñez-Martino, L. Fernández-Robles, E. Alegre, E. Fidalgo and I. Aguado-Caballero
A review of Towards Multi-Class Smishing Detection: A Novel Feature Vector Approach and the Smishing-4C Dataset 568
P. Blanco-Medina, A. Carofilis, E. Fidalgo, E. Alegre and I. Aguado-Caballero
Resumen de detección avanzada de smishing utilizando el dataset IRIS y visión artificial . 570
A. Martínez-Mendoza, L. Fernández-Robles, E. Alegre, E. Fidalgo and E. Díez-del-Río
A review of SpamClus: An agglomerative clustering algorithm for spam email campaigns detection 572
A. Gómez-Pascual, V. González-Castro, A. Martínez-Mendoza and C. Gaul
Comparative analysis of Natural Language Processing methods for Authorship Attribution of texts of different lengths 574
R. Valcarce-Álvarez, F. Jáñez-Martino, R. Alaiz-Rodríguez and V. González-Castro
A Comparative Study of Natural Language Processing Models for Detecting Obfuscated Text 579
A. Martínez-Mendoza, E. Alegre, E. Fidalgo, C. Gaul and I. Aguado-Caballero
A review of DeepHSAR: semi-supervised fine-grained learning for multi-label human sexual activity recognition 584
S. Muñoz Martín, L. A. Martínez Hernández, S. Pérez Arteaga, A. L. Sandoval Orozco and L. J. García Villalba
Detección de vulnerabilidades en TensorFlow comparando redes BERT y LSTM 586
L. Vigoya, C. Fernández Rosales, C. Alba Jiménez, R. Román and R. Morillas
RITA: Metodología Para Evaluación de Respuestas Inteligentes Aplicando Taxonomías de Alucinaciones 591

Sesión Pósters IV: Infraestructuras críticas, IoT y sistemas industriales 597

V. Machaka, S. Figueroa-Lorenzo, S. Arrizabalaga and J. Hernantes
A review of: "Standalone vs Hybrid SDN solutions for early detection of network channel attacks in Industrial Control Systems" . . . 597
S. Ruiz-Villafranca, J. Roldán-Gómez, J. M. Castelo Gómez, J. Carrillo-Mondéjar and J. L. Martínez
A Review of "A TabPFN-based intrusion detection system for the industrial internet of things" . . . 600
E. D. Gutiérrez Mlot, J. Saldana, R. J. Rodríguez, I. Kotsiuba and C. H. Gañan
A Review of "A dataset to train intrusion detection systems based on machine learning models for electrical substations" . . . 602
F. Bagheri-Gisour Marandyn, N. Rodríguez-Pérez, M. Castro Ponce, L. Sigrist and G. López López
Estimación Probabilística del Riesgo en Redes Eléctricas ante Ataques MaDIoT: Un Enfoque Multicapa . . . 604
M. Rivera-Dourado, M. Gestal, A. Pazos and J. Vázquez-Naya
A review of: A Novel Protocol Using Captive Portals for FIDO2 Network Authentication . 609
S. de Diego, C. Regueiro and G. Maciá-Fernández
A review of: Credenciales Colaborativas para el Internet de las Cosas . . . 611
M. T. Montero Ruiz and M. P. Marco Francia
El delito de estafa informática y el phishing tras la reforma de la LO 14/2022. La ciberseguridad como reto . . . 613
A. Baños Ramos, M. Reneses Botija, M. Castro Ponce, E. Awad and G. López López
Análisis y validación de un videojuego como herramienta de investigación para el ciberacoso 618
A. Gómez-Goiri, I. Gutierrez-Aguero and D. Garcia-Estevez
Una revisión de: Leveraging Blockchain Usage to Enhance Slag Exchange . . . 622
J. C. Ortiz-Pradillo and M. P. Marco Francia
Cooperación público-privada y protección de datos en la obtención transfronteriza de datos de localización como prueba electronica . . . 624

Premios RENIC a la mejor tesis doctoral y al mejor trabajo fin de máster 631

Premio al mejor trabajo fin de máster 631
J. Parada Tallante
TANUKIPOT: Honeypot basado en gemelos digitales para infraestructuras críticas . . . 631

Premio a la mejor tesis doctoral 641
D. Uroz Hinarejos
Advances in Cybersecurity Incident Prevention and Analysis . . . 641

Índice de Autores 647

Conferencias invitadas

Oracle Parfait: Real-World Vulnerability Detection at Scale and AI-Assisted Patch Generation

Ponente: Dr. Cristina Cifuentes, Oracle



Resumen

The Parfait static code analysis tool focuses on detecting vulnerabilities that matter in widely used languages including Java, C, C++, Go, Python, Swift and more. Its focus has been on key items expected out of a commercial tool, namely, precision of results (i.e., high true positive rate), scalability (i.e., being able to run quickly over millions of lines of code), incremental analysis (i.e., being able to run over deltas of the code quickly), and usability (i.e., ease of integration into standard build processes, reporting of traces to the vulnerable location, etc). Today, Oracle Parfait is used by thousands of developers at Oracle worldwide on a day-to-day basis. In this presentation we'll sample a flavour of Parfait — we explore some real world challenges faced in the creation of a robust vulnerability detection tool, look into various examples of vulnerabilities, and recount what matters to developers for integrating such a tool into today's continuous integration (CI) pipelines. I also provide a peek into the future – recent experiments on automatic generation of patches for vulnerabilities reported by Oracle Parfait through program and AI-assisted analyses.

Biografía

Cristina is the Vice President of the Oracle Software Assurance organisation where she leads a team of security researchers and software and machine learning engineers to make application security and software assurance, at scale, a reality. She was the founding Director of Oracle Labs Australia in 2010, where she led a team of researchers and engineers for close to 12 years, with a focus on scaling up Program Analysis techniques in new application security tools. Cristina led and successfully released Oracle Parfait, a static analysis tool used by thousands of C, C++ and Java developers each day. Cristina's passion for tackling the big issues in the field of Program Analysis began with her PhD work in binary decompilation at the Queensland University of Technology, which led to her being named the Mother of Decompilation for her pioneering contributions to this domain.

¿Cómo atacan los que atacan en el 2025?

David Sancho, Trend Micro



Resumen

En esta charla se explicará el panorama de amenazas actual haciendo hincapié en los nuevos métodos, como la inteligencia artificial. Las tendencias criminales en materia de ciberataques han ido evolucionando poco a poco introduciendo nuevas piezas y mejorando las antiguas. Se tocará la evolución reciente del ransomware y se contará cuáles son las mayores amenazas basadas en la IA en el momento actual. Tocaremos el tema deepfakes y amenazas a la infraestructura de la IA, entre otras cosas.

Biografía

David Sancho joined Trend Micro in 2002, having fulfilled a variety of technical security-related roles. Currently, his title is Senior Threat Researcher, and he specializes in web threats and other emerging technologies. In his more than 27 years of experience in the security field, David has written and published a number of research papers on malware tendencies, has been featured in the media, and has participated in customer events where he has presented on business issues and malware-related topics. His interests include web infection methods, vulnerability exploitation, and white-hat hacking in general. He also lifts weights. Sometimes

Programa científico

Sesión Investigación I: Privacidad, regulación y transparencia

Dataset de Licencia y Términos de Uso de Apps Móviles y Análisis de Riesgos de Privacidad

Enrique García-Arias*, Antonio Robles-Gómez*, Rafael Pastor-Vargas*, Llanos Tobarra*, Pedro Vidal-Balboa†, Mariano Méndez-Suárez‡

*Departamento de Sistemas de Comunicación y Control, ETSI Informática
Universidad Nacional de Educación a Distancia (UNED)
Madrid, España
Email: {e.garcia,arobles,rpastor,llanos}@scc.uned.es

†Consorcio CA UNED en Ponferrada / INTECCA
León, España
Email: pvidal@intecca.uned.es

‡ESIC University, ESIC Business & Marketing School
Madrid, España
Email: mariano.mendez@esic.university

***Resumen*—La implementación de normativas como el RGPD en Europa o la LPCC en USA ha obligado a los desarrolladores de aplicaciones al cumplimiento de criterios ligados a los derechos de privacidad de los usuarios. Este cumplimiento se refleja en unas políticas de privacidad y términos de servicio que exponen al usuario el uso de sus datos personales y requieren su aceptación para utilizar la aplicación. Sin embargo, este criterio les enfrenta ante la complejidad para la interpretación de estos extensos textos legales y la dificultad para comprobar que reflejen adecuadamente las funciones desarrolladas en el código de la aplicación. Para abordar estos retos, presentamos un análisis de extremo a extremo, desde la normativa de privacidad hasta los permisos solicitados y realmente utilizados en el código de la aplicación, haciendo un uso eficiente de las tecnologías de IA disponibles. Los resultados de esta investigación tienen actualmente un carácter privado como paso previo a su puesta en disponibilidad en abierto.**

***Keywords*—** Aplicaciones Móviles, Riesgos de Privacidad, Reglamento General de Protección de Datos (RGPD), Inteligencia Artificial (IA)

I. Introducción

El creciente desarrollo de las funcionalidades de las aplicaciones móviles presenta desafíos a las garantías de privacidad de los usuarios debido a la recopilación y el uso potencialmente indebido de sus datos personales. Las políticas de privacidad, que son el documento que debe describir el uso de estos datos por parte de las aplicaciones, adolecen de falta de transparencia que, junto con la complejidad para comprender los permisos otorgados a las aplicaciones, pueden facilitar la explotación de información sensible, el rastreo no autorizado y la generación de perfiles detallados con fines comerciales, o incluso maliciosos.

Estos riesgos se intensifican en poblaciones vulnerables y en el contexto de las redes sociales, donde la exposición de información sensible o de carácter personal sin un consentimiento informado adecuado, puede derivar en consecuencias adversas. En este escenario, es fundamental el desarrollo de mecanismos de evaluación automatizados y objetivos que permitan identificar estas vulnerabilidades de manera sistemática, mitigando riesgos y promoviendo entornos digitales más seguros y alineados con la normativa vigente.

En las últimas dos décadas, el marco regulador torno a la privacidad ha sufrido transformaciones significativas, destacando la implementación del Reglamento General de Protección de Datos (RGPD) en Europa y la Ley de Privacidad del Consumidor de California (LPCC). Estas regulaciones buscan otorgar a los usuarios un mayor control sobre sus datos personales, estableciendo requisitos específicos para su recopilación, almacenamiento y procesamiento. Sin embargo, el cumplimiento de estas normativas por parte de los desarrolladores de aplicaciones sigue siendo un desafío, pues es necesario garantizar la coherencia entre las políticas de privacidad y las prácticas reales de manejo de datos en el código de las aplicaciones.

Las políticas de privacidad son el instrumento que refleja el compromiso entre desarrolladores y usuarios. No obstante, el usuario se ve superado por la complejidad y extensión de los textos legales componentes de las políticas de privacidad y términos de servicio que acepta sin lectura. Y, por otra parte, el usuario debe presuponer que las funciones y permisos utilizados realmente por el código de la aplicación están expresados adecuadamente en las políticas de privacidad.

Al igual que cualquier documento legal, las políticas de privacidad están sujetas a interpretación. El principal desafío radica en la naturaleza del lenguaje jurídico utilizado y la variación que se produce entre diferentes sectores empresariales, donde la experiencia en un sector puede no ser aplicable a otro. Para los usuarios, la complejidad de estos documentos, lleva a evitar su lectura debido al tiempo requerido y al nivel académico necesario para su comprensión [1], [2]. Las aplicaciones o sitios web enlazan a la política de privacidad de su organización, que suele ser bastante genérica al estar diseñada para aplicarse a todos los productos y servicios de la empresa. Esto presenta retos relacionados con la legibilidad [3], [4], [5], la ambigüedad [6], y la accesibilidad de la información [7], [8]. Con la implementación del RGPD, las políticas de privacidad han sido actualizadas, incrementando su extensión y detalle [9], [10].

El continuo dinamismo en el desarrollo de nuevas aplicaciones ha generado textos de políticas de privacidad sin un formato coherente y con una notable inconsistencia en relación con el flujo de tratamiento de datos dentro del software. Muchas aplicaciones otorgan privilegios excesivos debido a una inadecuada correlación entre sus políticas de privacidad y el marco regulatorio vigente. Estas discrepancias incluyen el acceso a datos confidenciales a través de permisos peligrosos, el uso de bibliotecas de terceros, la difusión de datos, la interacción con APIs y otros elementos relacionados, como el código de la aplicación, la inclusión de metadatos y las prácticas de privacidad [9], [11], [12], [13], [14], [15], [16].

El objetivo de esta investigación es dotar al usuario de las aplica-

ciones móviles de instrumentos objetivos que le permitan identificar los riesgos de privacidad que pueden derivarse de su instalación y uso. Para conseguir este objetivo, respondemos a nuestra pregunta de investigación: "¿Cómo identificar la coherencia entre los requerimientos normativos de privacidad, los permisos declarados y los permisos realmente solicitados por las aplicaciones, y cómo pueden mejorarse los métodos de detección utilizando Deep Learning y herramientas de análisis de código?". Los resultados obtenidos se orientan a empoderar al usuario con instrumentos objetivos que le permiten identificar de forma intuitiva y comprensible los riesgos de privacidad y el nivel de cumplimiento de la normativa. De forma complementaria, permiten a los proveedores de servicio y desarrolladores de aplicaciones mejorar sus políticas y prácticas de privacidad hacia los usuarios.

El resto de esta contribución está organizada de la siguiente manera. La Sección II aborda en primer lugar un análisis de los estudios previos existentes para identificar fortalezas como las metodologías de creación de corpus y de cumplimiento de las normativas de privacidad, así como limitaciones entre las que destaca la carencia de un análisis automatizado e integral de los riesgos, que son motivación de este trabajo. En la Sección III se detallan las fases involucradas en el desarrollo de las funcionalidades base abordadas en la propuesta de investigación y desarrolladas en cada una de ellas. La Sección IV expone los criterios que dirigen el desarrollo, incluyendo la calidad de los datos del corpus, el enfoque de cumplimiento de la normativa y análisis de permisos, la funcionalidad de integración del conjunto del proceso. También tratamos la adecuación de las herramientas de Deep Learning para la sistematización de los criterios de generación de los Datasets que son base de nuestros resultados. Por su parte, en la Sección V describimos las métricas objetivas, resultado del proceso y que fundamentan la evaluación de los riesgos de privacidad. Finalmente, en la Sección VI se concluye este trabajo y se proponen posibles trabajos futuros.

II. Estudios Previos

Los estudios más relevantes en nuestro ámbito de investigación han abordado distintos enfoques en torno a la transparencia y cumplimiento normativo de las aplicaciones móviles. Por un lado, una línea de investigación se ha centrado en la mejora de la comprensión y legibilidad donde destacan varios proyectos orientados a mejorar la legibilidad de las políticas de privacidad, permitiendo a los usuarios alinear sus preferencias de privacidad con las políticas ofrecidas por las aplicaciones. PolicyLR [17] convierte las políticas de privacidad en un formato legible por máquina utilizando lógica formal y "atomic formulae". Esto posibilita la definición formal de tareas como el cumplimiento y la consistencia.

A su vez las técnicas de clasificación de textos han sido útiles para permitir a los usuarios obtener una visión general de alto nivel de una política a través de etiquetas de categorías predefinidas [18], [[19] mediante el uso de corpus anotados. Un ejemplo es APPcorp [20] que ha creado un corpus etiquetado con 231 políticas de privacidad, lo que facilita el análisis automático de las estructuras documentales de estas políticas, contribuyendo a una mejor comprensión para los usuarios.

Sin embargo, persisten los desafíos en la interpretación y análisis de las políticas de privacidad [21]. PoliGraph [12] aborda la incompletitud semántica mediante un enfoque basado en grafos de conocimiento que analiza políticas completas utilizando el etiquetado de roles semánticos. De igual manera, los enfoques basados en ontologías ofrecen capacidades avanzadas para el análisis semántico de políticas de privacidad. Proyectos como OPPO [19], Pronto [22] y PrivOnto [23], aprovechan estas tecnologías para representar y analizar las políticas, permitiendo consultas semánticas más detalladas.

Otros estudios han abordado la coherencia entre las políticas declaradas y el comportamiento real de las aplicaciones. En este sentido, uno de los principales objetivos ha sido identificar las discrepancias entre lo estipulado en los documentos legales y la recopilación efectiva de datos personales a través de permisos [24]. Mediante el análisis de la aplicación, es posible convertir los permisos, las APIs y la interfaz gráfica de usuario en un conjunto de información personal, que se recopila para evaluar la coherencia entre el comportamiento de la aplicación y la política de privacidad establecida.

Este proceso debe centrarse en los datos considerados sensibles o "peligrosos". Para determinar si la aplicación mantiene coherencia en el uso de permisos, PermPress [8], utilizando el corpus MPP-270, analiza los permisos peligrosos declarados en el archivo de Manifiesto (manifest.xml). Además, otro uso del corpus MPP-270 es la generación automatizada de los textos de política de privacidad, lo que garantiza el cumplimiento de la normativa RGPD durante el diseño de la aplicación mediante el uso de diagramas UML.

En el ámbito del cumplimiento normativo, los esfuerzos se han concentrado en evaluar el grado de conformidad de las aplicaciones con los requisitos legales establecidos en el RGPD. Los métodos automatizados para evaluar el cumplimiento de las políticas con las regulaciones son un área de creciente interés. En esta línea el framework ARC [25] transforma las complejas regulaciones de privacidad en un formato estructurado, facilitando el análisis automatizado y la evaluación del cumplimiento. PolicyChecker [26] emplea un enfoque basado en el etiquetado de reglas y roles semánticos para evaluar el cumplimiento de las políticas de privacidad de aplicaciones móviles con el RGPD. PTPDroid [5] identifica posibles infracciones relacionadas con las prácticas de recopilación de datos por parte de terceros, mientras que Polisis [21] analiza el impacto del RGPD en las políticas de privacidad. Además, la generación de "privacy policy captions" mejora la transparencia para el usuario al ser precisas, consistentes y fáciles de entender, reflejando el grado de cumplimiento con los requisitos reglamentarios. El sistema Knight [10] se centra en la asociación de los textos de las políticas de privacidad con artículos específicos de RGPD, Pregunta/Respuesta, que permiten realizar consultas de manera libre. El artículo 13 del RGPD, que detalla la información que los proveedores y fabricantes deben proporcionar cuando recopilan datos personales, es un foco de atención en diversas investigaciones.

En términos de usabilidad, diversos proyectos han desarrollado soluciones que facilitan a los usuarios la comprensión y gestión de su privacidad. Funcionalidades basadas en modelos de pregunta-respuesta permiten realizar consultas de manera libre, entre los proyectos más relevantes se encuentran PriBot [21], la creación del corpus PrivacyQA [27], y el dataset PolicyQA [28]. Estos avances amplían la funcionalidad de clasificación y respuesta a preguntas, apoyando nuevas investigaciones.

El proyecto de Usable Privacy Policy (UPPP) ha desempeñado un papel clave en la creación de corpus y el desarrollo de técnicas de análisis y herramientas que ayudan a los usuarios a comprender y gestionar mejor su privacidad. Un ejemplo de esto es la utilización de las funcionalidades de Pregunta/Respuesta [6]. Opt-Out [29] facilita la extracción automática de enlaces de exclusión en las políticas de privacidad [30], simplificando así la toma de decisiones por parte de los usuarios.

Otro campo de interés es la generación de resúmenes (summarization), que destaca la información clara y relevante para el usuario. Los proyectos se basan en la aplicación de técnicas de Procesamiento del Lenguaje Natural (PLN). Los autores de [6] amplían esta funcionalidad, permitiendo realizar análisis comparativos entre aplicaciones competidoras. En [7] se presenta el primer corpus bilingüe de políticas de privacidad de aplicaciones móviles, compuesto por 64 políticas en inglés y 91 en alemán, con anotaciones manuales.

Sin embargo, a pesar de los avances en estas áreas, aún persisten diversas limitaciones en los enfoques actuales. En general se evidencia la carencia de un análisis multi-nivel integrando la normativa RGPD, las políticas de privacidad, los términos de servicio y el código fuente. Los trabajos que abordan la clasificación en categorías, no ofrecen una evaluación global de los riesgos y aquellos que realizan un análisis del código no relacionan sus resultados con las implicaciones de riesgos de privacidad señaladas por la normativa y refrendados por el compromiso del desarrollador en las políticas de privacidad y términos de servicio. Otra limitación es que la mayoría de las investigaciones se basan en textos en inglés, existiendo una notable carencia de estudios bilingües, especialmente en castellano.

En este contexto, el trabajo busca superar estas limitaciones mediante un enfoque integral que combine el análisis normativo,

documental y técnico, garantizando una evaluación más objetiva y automatizada de la privacidad en aplicaciones móviles.

III. Proceso de Generación del Dataset

El proceso de generación automático de nuestro Dataset, e identificación y análisis de riesgos de privacidad en aplicaciones móviles que rigen la metodología de investigación de este trabajo se estructura en 5 fases representadas en la Figura 1. La Fase I consiste en la creación del corpus incluyendo las políticas de privacidad, los términos de servicio y el código de aplicación. La Fase II analiza las categorías relacionadas con la normativa RGPD. La Fase III valora su grado de cumplimiento. La Fase IV identifica las categorías de permisos críticos en las políticas de privacidad, y la Fase V hace un análisis en profundidad del código fuente de aplicación identificando el uso de los permisos críticos. A continuación se van a ir describiendo con detalle cada una de estas cinco fases.

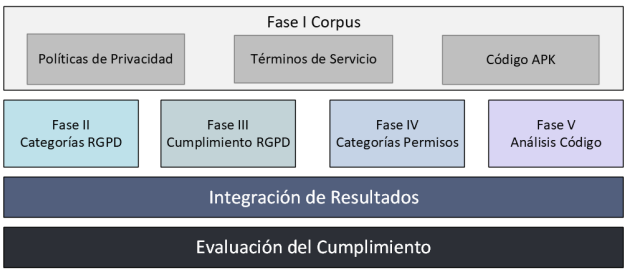


Figura 1: Esquema Global

III-A. Creación del Corpus

El objetivo de esta fase es construir un corpus estructurado de las aplicaciones que recopile de forma estructurada la Política de Privacidad, los Términos de Servicio, y el código de aplicación.

En el marco del proyecto "Análisis de aplicaciones móviles desde la perspectiva de la protección de datos: Ciberprotección y ciberriesgos de la información de la ciudadanía", cuyo foco se centra en aquellas aplicaciones con mayor impacto previsible sobre la privacidad, se ha delimitado el área de estudio a las aplicaciones de redes sociales, al ser estas, sin lugar a dudas, las más utilizadas a nivel global. Por ello, el corpus ´ se ha restringido 200 aplicaciones de redes sociales con mayor número de descargas de Google Play[31]. El código de aplicación se almacena en formato APK procediendo a su descompilación, mientras que las Políticas de Privacidad y los Términos de Servicio se recopilan mediante procesos automatizados de rastreo (Crawlers). Estos procesos descargan los documentos originales y convierten los formatos HTML a texto plano realizando los necesarios tratamientos de limpieza y preproceso (Data Cleaning y Data Curation), para componer un Corpus actualizado que incluye Políticas de Privacidad y Términos de Servicio en formato bilingüe inglés y castellano.

La complejidad del proceso deriva del automatismo necesario para acceder a los documentos originales que contemplan formatos distintos y uso de facilidades JavaScript. Adicionalmente, así como las políticas de privacidad están referenciadas en la mayoría de casos en el archivo Manifest del código de aplicación, no sucede lo mismo con los términos de servicio por lo que debe añadirse un proceso específico de búsqueda en estos casos. El requerimiento de disponer de los textos en lengua inglesa así como en castellana cuando esté disponible añade más complejidad al tratamiento y es seguramente una de las razones de la carencia de estudios similares que incluyan en el corpus tanto los términos de servicio como los documentos de políticas de privacidad y términos de servicio en formato bilingüe.

A estas complejidades debemos añadir que el objetivo es que sea un proceso actualizado de modo que contempla las actualizaciones de los documentos y mantiene una gestión de ejecuciones y actualizaciones de forma automatizada. Este aspecto es necesario si se desea mantener la vigencia de los resultados de la investigación afrontando su obsolescencia y por tanto su pertinencia.

III-B. Categorías RGPD

En esta fase se identifican y categorizan los párrafos específicos de las políticas de privacidad en relación con la normativa RGPD. Como base del criterio de categorización, se ha tomado como referencia el corpus etiquetado manualmente APPCorp [20] que contiene 231 políticas de privacidad (de más de 566 mil palabras y 7,748 párrafos anotados). Se propone un esquema de 11 etiquetas (categorías) directamente relacionadas con artículos del RGPD:

- Introducción de la política (RGPD Art.13.1.);
- Recolección y uso de primera parte;
- Cookies y tecnologías similares;
- Compartición con terceras partes (RGPD Art.13.2 b-f);
- Derechos y control del usuario (RGPD Art.32.1.);
- Seguridad de datos (RGPD Art.13.2. a);
- Retención de datos (RGPD Art.13.1. f);
- Transferencia internacional de datos (RGPD Art.40.2 g);
- Audiencias especiales;
- Actualización de la política (RGPD Art.13.1 a);
- Información de contacto de la política;
- Información de contacto del controlador de datos (primera parte).

Este proceso se realiza aplicando los métodos de Deep Learning descritos más adelante, el modelo de inferencia seleccionado en base a las métricas de precisión, garantiza la aplicación homogénea de las categorías en cada documento de política de privacidad descargado de las aplicaciones.

III-C. Cumplimiento RGPD

En esta fase se aplican los métodos desarrollados en PolicyChecker [26] que presenta un análisis automatizado a gran escala de 205973 políticas de privacidad de aplicaciones móviles para detectar violaciones de cumplimiento del RGPD.

Este trabajo utiliza varias tecnologías centradas en PLN y el análisis automatizado de textos legales. Las principales tecnologías empleadas son:

- Semantic Role Labeling (SRL). Basado en AllenNLP, para identificar estructuras semánticas en las oraciones y determinar quién realiza qué acción sobre qué datos.
- Named Entity Recognition (NER). Utilizado para identificar entidades clave como datos personales, controladores y usuarios. Se basa en un modelo adaptado de PurPliance.
- Modelos de Lenguaje Sentence-BERT (SBERT all-MiniLM-L6-v2). Empleado para evaluar la similitud semántica entre oraciones y plantillas predefinidas, orientadas a la identificación de términos relevantes del RGPD.
- Técnicas de Análisis Basadas en Reglas. Utilización de reglas para clasificar verbos y estructuras gramaticales presentes en las políticas de privacidad, para facilitar la identificación de prácticas de recolección, uso y transferencia de datos.
- Algoritmos de Lógicas condicionales. Utilizadas para identificar requisitos del RGPD que son obligatorios o aplicables bajo ciertas condiciones.

III-D. Categorías Permisos

Esta fase se basa en la identificación de menciones de permisos en las políticas de privacidad. Tras una primera evaluación de los corpus OPP-115 y MPP270 [8], identificamos que la mayoría de las investigaciones disponibles realizan un análisis sobre los grupos de permisos. Para realizar una correcta correlación de los permisos mencionados en las políticas de privacidad con los permisos realmente declarados en el código es necesaria una mayor granularidad en la identificación de estos permisos.

Para abordar estas limitaciones, se han adoptado estrategias basadas en Modelos LLM (Large Language Models), en particular LLAMA 2 (Meta-Llama-3-8B-Instruct/original/). Los LLMs, como ChatGPT y Llama 2, han demostrado un alto rendimiento en el procesamiento de lenguaje complejo y en la identificación de patrones en grandes volúmenes de texto, lo que los hace especialmente adecuados para extraer prácticas clave de las políticas de privacidad [32].

Con este enfoque se consigue categorizar los párrafos de las Políticas de Privacidad con los permisos considerados peligrosos, tal

Grupo	Permiso
CALENDAR	READ_CALENDAR, WRITE_CALENDAR
CAMERA	
CONTACTS	READ_CONTACTS, WRITE_CONTACTS, GET_ACCOUNTS
LOCATION	ACCESS_FINE_LOCATION, ACCESS_COARSE_LOCATION
MICROPHONE	RECORD_AUDIO
PHONE	READ_PHONE_STATE, CALL_PHONE, READ_CALL_LOG, WRITE_CALL_LOG, ADD_VOICEMAIL, USE_SIP, ANSWER_PHONE_CALLS
BODY_SENSORS	-
SMS	SEND_SMS, RECEIVE_SMS, READ_SMS, RECEIVE_WAP_PUSH, RECEIVE_MMS
STORAGE	READ_EXTERNAL_STORAGE, WRITE_EXTERNAL_STORAGE

Tabla I: Clasificación de Permisos por Grupo

y como se puede apreciar en la Tabla I. Estos permisos corresponden a los identificados en código en la Fase V Análisis de Código, lo que facilita relacionar los textos de las políticas de Privacidad con los utilizados por la aplicación en el código.

III-E. Análisis de Código

En esta fase se hace un uso extensivo de la herramienta Androguard realizando un proceso de ingeniería inversa para extraer la información de la estructura y comportamiento de la aplicación. De este modo, podemos identificar los permisos realmente utilizados en el código de una aplicación, en contraste con los declarados en su manifiesto, permitiendo detectar discrepancias que podrían implicar riesgos de privacidad. A través del análisis del bytecode y la identificación de clases, métodos y llamadas a APIs, generamos los datos necesarios para la compilación de un dataset estructurado que relaciona los permisos empleados con los componentes específicos del código, proporcionando una base sólida para la evaluación automatizada de riesgos de seguridad y privacidad.

IV. Metodología de Investigación

La metodología propuesta se ha basado en los criterios de ser repetible, permitiendo la automatización del análisis; objetiva, basada en datos y procesos verificables; medible, al generar resultados cuantificables; integral, abordando el cumplimiento de la privacidad desde distintas dimensiones; rigurosa, aplicando estándares sólidos y basada en la construcción de un corpus completo y actualizado; integrada, todos los componentes y resultados de los procesos se interrelacionan en una arquitectura común y forman parte de un proyecto amplio de privacidad denominado Safe Mountain [33].

IV-A. Corpus Base del Proceso

La calidad, actualización y precisión de los datos del corpus base es crítico en este estudio. Se han seleccionado 200 aplicaciones en el contexto de las redes sociales, de las cuales se extraen de manera automatizada sus Políticas de Privacidad, Términos de Servicio, archivos APK y metadatos asociados. Este enfoque automatizado garantiza la coherencia y consistencia de los datos, y previene el uso de información obsoleta, asegurando que los análisis obtengan resultados actualizados.

Dado el volumen de información procesada, que abarca documentos en inglés y castellano, se requiere un pipeline optimizado para manejar eficientemente la extracción y actualización de estos datos. La automatización de este proceso es fundamental, ya que la magnitud del corpus hace inviable una gestión manual, cualquier desajuste en la actualización de los datos comprometería la pertinencia de los posteriores análisis.

El corpus no es un objetivo final en sí mismo, sino la base fundamental para el desarrollo de un análisis más amplio. A diferencia de otros estudios que se limitan a la construcción de corpus de políticas de privacidad, nuestro enfoque trasciende esta fase inicial y utiliza estos datos como fuente para la aplicación de modelos avanzados de Deep Learning donde la fiabilidad y actualización de los datos resultan determinantes para la precisión de los modelos de PLN.

Por lo tanto, este trabajo no solo pone énfasis en la recopilación de datos, sino en su mantenimiento continuo y su integración dentro de un sistema automatizado y escalable, asegurando que cada componente del análisis de privacidad cuente con información precisa y actualizada.

IV-B. Cumplimiento RGPD

Para el análisis del cumplimiento de la normativa RGPD, se utilizan corpus anotados manualmente y ampliamente validados en los estudios previos, incluyendo APPCorp [20] y OPP115 [34]. Partiendo de estos datos se aplican las reglas de Deep Learning, complementadas con técnicas basadas en reglas y etiquetado semántico (SRL, NER, etc.).

Los resultados del proceso se organizan en métricas que evalúan el grado de cumplimiento de la normativa RGPD, identifican los requerimientos ambiguos y analizan la coherencia general de las políticas de privacidad. Además, a través de los procesos de categorización, se establece una relación estructurada entre los elementos textuales mencionados en las Políticas de Privacidad, permitiendo la trazabilidad de un análisis más preciso y detallado.

IV-C. Análisis de Permisos

Este proceso consiste en la categorización y mapeo de los permisos individuales con los grupos de permisos críticos. A través del análisis exhaustivo del código de aplicación se correlacionan los permisos declarados con el comportamiento real de la aplicación.

El uso de Deep Learning y en concreto de modelos LLM permite una categorización más eficiente y detallada, facilitando la detección de incoherencias y posibles vulnerabilidades con un alto nivel de granularidad. En este apartado se combinan las técnicas de Proceso de lenguaje Natural con aquellas más orientadas al análisis en profundidad del código, relacionando componentes intrínsecos de la programación con sentencias de textos de naturaleza legal.

IV-D. Integración

El componente de integración es fundamental para la consecución del objetivo de un análisis de extremo a extremo de los riesgos de privacidad. El enfoque en fases permite abordar de forma especializada cada componente que tiene influencia en la evaluación de los criterios de privacidad, gran parte de estos desarrollos especializados se benefician del conocimiento de estudios previos que abordan aspectos individuales de la privacidad, añadiendo las necesarias contribuciones y análisis aportados por este estudio. Sin embargo nuestra aportación esencial es la de diseñar un ecosistema que capitaliza los resultados de cada fase poniéndolos en relación con nuestro objetivo global definido en la Sección I.

Cada fase se alimenta de los corpus generados y mediante la sistematización generada por los algoritmos de Deep Learning genera los correspondientes Datasets que son el fundamento de la elaboración de las métricas globales de privacidad que reflejan el resultado de la integración de los diferentes procesos especializados descritos anteriormente.

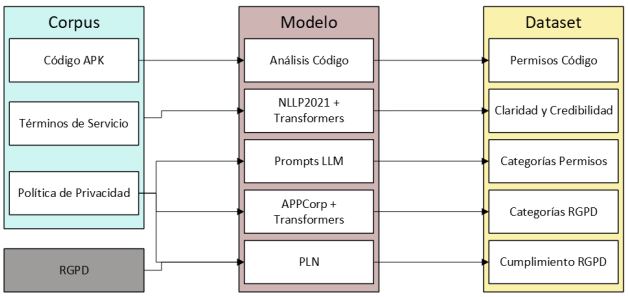


Figura 2: Generación de Datasets

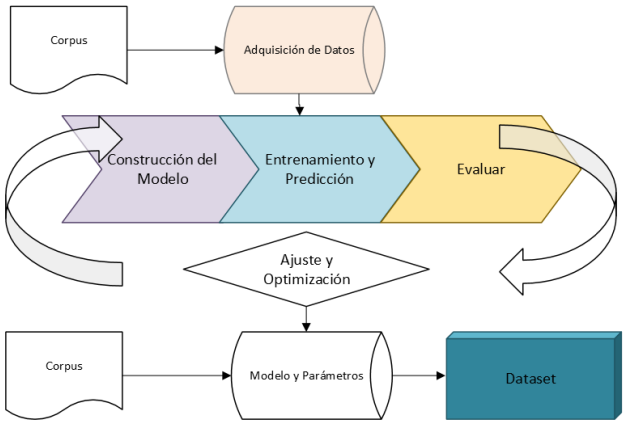


Figura 3: Ejecución Deep Learning

La Figura 2 representa el proceso que parte de la información estructurada en el Corpus de aplicaciones y culmina en la generación de los Datasets estructurados.

IV-E. Deep Learning

En cada una de las fases anteriores es preciso la utilización de las tecnologías de Deep Learning, además de técnicas de PLN, para la identificación y categorización de los indicadores de privacidad. Los procesos de Deep Learning sistematizan la información textual que de esta forma puede ser procesada para componer una visión unificada del cumplimiento de privacidad por parte de las aplicaciones. En la Figura 2 los procesos de Deep Learning se corresponden a la columna "Modelo" y los resultados componen los correspondientes "Datasets" que pueden ser puestos en relación para una evaluación objetiva del cumplimiento de privacidad.

La Figura 3 representa los procesos realizados con Deep Learning, desde la incorporación de los datos de los Corpus Anotados, al aprendizaje del modelo y su aplicación para inferencia de las categorías en el Corpus de Políticas de Privacidad y Términos de Servicio. En el módulo de *Adquisición de Datos* se aborda la complejidad derivada de la diversidad de los datos recopilados en los corpus dado que afecta directamente a los resultados de los modelos de Deep Learning. Estos datos están limitados por el costo de los procesos de anotación manual, que requieren que los investigadores analicen minuciosamente las políticas de privacidad. Estas limitaciones, sin el adecuado tratamiento, resultan en que muchos textos no se asocian a las categorizaciones adecuadas.

Para abordar estas complejidades, este componente incluye limpieza de texto, tokenización, lematización, stemming, stopwords, padding y data augmentation. La codificación contempla:

- Embeddings. Representaciones densas preentrenadas que capturan relaciones semánticas.
- TextRepresentations. Conversión del texto en vectores basados en la frecuencia de las palabras o caracteres.

En los procesos de entrenamiento e inferencia, correspondientes a los módulos de *Construcción del Modelo, Entrenamiento y Predicción y Evaluación* de la Figura 3, se han utilizado diferentes algoritmos incluyendo redes neuronales (RNN, ByLSTM, GRU) y transformers (BERT, RoBERTa, DistilBERT, AlBERT) utilizando una arquitectura común que permita comparar los resultados.

Esta arquitectura facilita la asignación de hiperparámetros de forma flexible lo que permite optimizar los resultados de forma incremental y controlada reflejado en la Figura 3 con *Ajuste y Optimización*, que sintetizamos en 3 fases:

- Fase inicial. Se establecen valores básicos como son: tamaño del vocabulario, número de épocas, tamaño del batch, proporción de los datos de validación, "data augmentation" o "dropout".
- Fase de mejora. Ajustes de los valores iniciales de: tasa de Aprendizaje, tamaño del lote (Batch Size), gestión del sobreajuste, escalado de dimensiones de salida y dimensiones ocultas, congelación de capas en "Embeddings", entre otros.
- Fase avanzada. Gestión de los gradientes, procesos de acumulación para gestión de memoria, sobremuestreo y balanceo de clases, normalización de capas y "pooling" y ajustes específicos por algoritmo.

Para la mejora de la detección de las clases menos frecuentes y reducir el excesivo impacto de las más representadas, se han aplicado estrategias de aumento de datos (sustitución por sinónimos, back-translation, random insertion, random swap o random deletion), attention mask, smote (Synthetic Minority Over-sampling Technique) y class balance entre otros.

Para el proceso de aprendizaje han sido relevantes la definición del tamaño del lote (BatchSize) dado que los lotes grandes tienden a estabilizar el entrenamiento, pero pueden reducir ligeramente la precisión, ajustar el dropout evitando el sobreajuste, utilizar early-stopping (detención temprana) para evitar el sobreajuste, congelación de capas iniciales en los modelos transformers.

IV-F. Gestionar el proceso y evaluar el rendimiento

La evaluación y mejora de los resultados se apoya en una primera fase en los indicadores de: Exactitud (Accuracy): porcentaje de predicciones correctas sobre el total; Precision (Precision): porcentaje de verdaderos positivos entre los ejemplos clasificados como positivos; y Matriz de Confusión (Confusion Matrix): tabla que muestra las verdaderas predicciones contra las predicciones fallidas. Otros indicadores provistos para mayor detalle son Exhaustividad (Recall), Índice F1 (F1 Score), Índice F2 (F2 Score), Coeficiente de Correlación de Matthews (CCN) (Matthews Coheficient Correlation), Exactitud Balanceada (Balanced Accuracy), Cohen Kappa, Log de Pérdida (Log Loss) y Clases Predichas (Classes Predicted).

Tal como hemos indicado anteriormente las pruebas se han realizado con una amplia gama de algoritmos, a continuación, en la Tabla II, mostramos un extracto de estos indicadores aplicado a los algoritmos de Deep Learning que han mostrado mejores resultados aplicados a los distintos corpus de entrenamiento que corresponden a las distintas fases mencionadas.

Algoritmo	Corpus	Exactitud	Precisión	Exhaustividad
ByLSTM	NLLP2021	90.54 %	91.32 %	90.54 %
DistilBERT	NLLP2021	95.79 %	95.96 %	95.79 %
RoBERTa	MPP270	92.51 %	93.08 %	92.51 %
DistilBERT	MPP270	94.25 %	94.62 %	94.25 %
ByLSTM	MPP270	93.83 %	94.77 %	93.83 %
DistilBERT	APPCorp	95.09 %	95.18 %	95.09 %
ByLSTM	APPCorp	94.83 %	94.98 %	94.83 %

Tabla II: Resultados de Evaluación de Algoritmos en Diferentes Corpus

En la aplicación de los modelos vemos que se obtienen unos buenos resultados muy influenciados por el proceso de ajuste requerido en cada caso. En el caso de RoBERTA nos vemos limitados por la elevada exigencia computacional lo que complica la ejecución de pruebas con Datasets extensos, para mejorar su eficiencia es preciso realizar un laborioso proceso de ajuste. El modelo DistilBERT es menos exigente en términos de memoria facilitando los procesos iterativos de prueba y ajuste con mínima pérdida de precisión. La utilización de ByLSTM está apoyada por el uso de embeddings preentrenados, se están utilizando principalmente FastText, RoBERTa, TF-IDF y GloVe. ByLSTM ha demostrado ser una buena opción con el uso de Datasets preentrenados de menor tamaño.

Las Figuras 4 y 5 reflejan los resultados del Deep Learning aplicando el modelo DistilBERT para clasificar permisos peligrosos en las Políticas de Privacidad. Se muestran dos informaciones:

- Tendencias en el Rendimiento del Modelo:
  - Curva de Pérdida: Tanto la pérdida en la fase de entrenamiento como en la de evaluación disminuye eficientemente

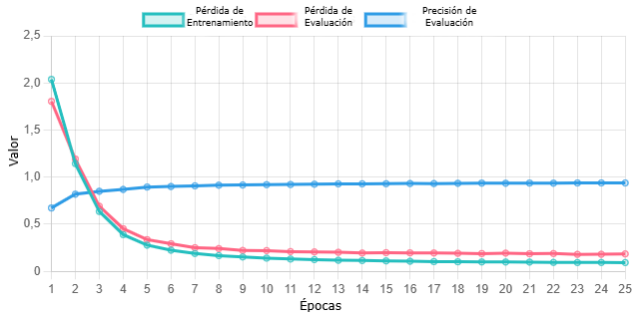


Figura 4: Curvas de Rendimiento del Modelo

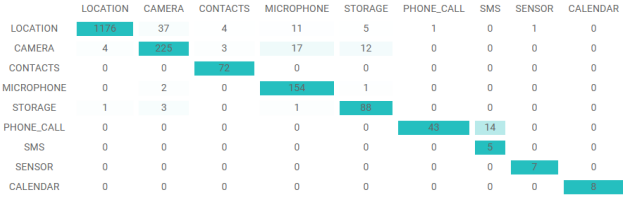


Figura 5: Matriz de Confusión

en las primeras épocas y se estabiliza alrededor de la época 10. Se evidencia un efectivo progreso de aprendizaje sin signos evidentes de sobreajuste.
- Curva de precisión: La precisión se estabiliza temprano, lo que indica que el modelo converge rápidamente, lo que resalta la eficiencia del modelo DistilBERT.

- Matriz de Confusión:
  - La categoría Ubicación obtiene la mayor cantidad de muestras clasificadas correctamente (1176), con una mínima cantidad de errores. Esta clase está sobrerepresentada en la muestra y ha sido necesario aplicar ajustes para el balanceo de clases en el proceso.
  - Las clases de cámara, micrófono y almacenamiento muestran cierta confusión posiblemente debido a similitudes contextuales en los textos.
  - Permisos como SMS, Sensor y Calendario tienen baja representación, lo que podría indicar una menor presencia en el corpus o dificultades del modelo para diferenciarlos.

Estos datos corresponden a un ejemplo concreto pero las tareas de mejora de resultados han debido aplicarse con cada modelo de categorización. Los resultados están muy influenciados por la calidad de los datos de partida; en este sentido, se eligió usar datos de entrenamiento anotados manualmente por expertos y cuyo coste de generación influyó en que las muestras sean de tamaño reducido o que contengan un significativo desbalanceo de clases.

En la Fase IV de Categorías de Permisos, se alcanzó una mayor eficiencia al emplear los modelos LLM, específicamente el modelo meta-llama/Meta-Llama-3-8B-Instruct, diseñado y optimizado para interacciones de tipo diálogo que requieren un alto grado de comprensión y respuesta contextual. El LLM no solo permite categorizar los permisos con la granularidad deseada, sino que también incluye una explicación detallada del criterio utilizado para cada asignación, lo que aporta transparencia y confianza al proceso de clasificación. La disponibilidad de un corpus bilingüe en inglés y español permite un análisis comparativo riguroso de las asignaciones de permisos entre las versiones inglesa y española de las políticas de privacidad de las aplicaciones. Estas diferencias, que reflejan variaciones en la interpretación o declaración de permisos según el idioma, se visualizan claramente en la Figura 6, destacando las discrepancias identificadas para cada aplicación.

Este enfoque no solo incrementa la utilidad del sistema para los desarrolladores y usuarios, sino que también permite nuevas áreas

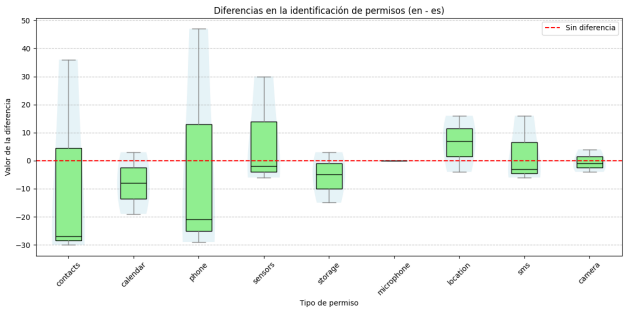


Figura 6: Comparativa de permisos En-Es

de investigación y aplicaciones en el ámbito de la privacidad y la seguridad de las aplicaciones móviles.

V. Aplicación de los Modelos: Indicadores y Métricas

Los procesos de análisis y categorización realizados sobre cada corpus asociado se lleva a cabo de manera automatizada mediante la aplicación de algoritmos y estrategias de Deep Learning. Como resultado de estos procesos, se generan los Datasets correspondientes, a partir de los cuales es posible identificar las siguientes métricas:

1. *Permisos críticos utilizados en código y no mencionados en la Política de privacidad.* Se identifican las posibles incoherencias entre el comportamiento real de la aplicación y la información proporcionada al usuario que pueden indicar que la privacidad del usuario no se protege adecuadamente o que la aplicación actúa ocultamente. Este indicador muestra falta de transparencia y posibles problemas de cumplimiento del marco regulatorio.
2. *Nivel y frecuencia de los permisos críticos utilizados.* Permisos críticos, como el acceso a la ubicación, la cámara o los contactos, pueden estar relacionados con la recolección de datos sensibles. El análisis del nivel de estos permisos y su frecuencia de uso permite identificar patrones negativos que afectan a la privacidad del usuario. El cumplimiento con las regulaciones de privacidad exige un uso mínimo y justificado de los datos sensibles.
3. *Nivel de claridad y equidad de los Términos de servicio.* Clasificando los párrafos de los Términos de Servicio como 'fair', 'unfair', 'potentially fair' o 'potentially unfair', este indicador permite identificar prácticas que podrían ser abusivas o poco transparentes. La clasificación inicial, realizada manualmente por un equipo experto en aspectos legales, asegura que los criterios de evaluación se basan en un conocimiento profundo de las normativas y derechos del usuario. La aplicación de un modelo de Deep Learning para automatizar la evaluación de nuevos textos, garantiza una supervisión continua y efectiva basada en los criterios mencionados.
4. *Nivel de cumplimiento de las directrices RGPD en la Política de Privacidad.* La evaluación de violaciones del cumplimiento de las directrices se realiza sobre 26 requerimientos RGPD identificados en la Fase III Cumplimiento RGPD y que cubren aspectos fundamentales del tratamiento de datos personales, como la identidad del controlador, los derechos de los usuarios (acceso, rectificación, borrado, etc.), la retención de datos, la base legal para el procesamiento, las transferencias internacionales de datos, y las decisiones automatizadas, entre otros. Un alto nivel de cumplimiento refleja un compromiso con la transparencia, la seguridad y la gestión adecuada de los datos personales.
5. *Nivel de actualización y coherencia con la actualización del código APK y la normativa.* La frecuencia de actualización en relación con el código es un indicador que, en conjunción con el resto, indica transparencia y compromiso con la garantía de calidad de la documentación y el alineamiento con la normativa.

El alcance de estas métricas permite revelar el grado de transparencia y cumplimiento normativo de las aplicaciones. Las incoherencias entre permisos declarados y utilizados evidencia la falta de transparencia y posibles incumplimientos, el nivel y frecuencia de los permisos críticos revela patrones de uso de datos sensibles, la clasificación automatizada de los términos de servicio en categorías de equidad permite detectar cláusulas potencialmente abusivas y el análisis del grado de cumplimiento de las directrices del RGPD verifica el nivel de alineación de las políticas de privacidad con los requerimientos legales.

La integración de estos indicadores en un proceso automatizado mejora la eficiencia y facilita el análisis y la supervisión continua de las aplicaciones, contribuyendo a una detección temprana de riesgos y la mejora de la seguridad y privacidad del usuario.

VI. Conclusiones Y Trabajos Futuros

En esta investigación hemos descrito un proceso automatizado y transparente para el diagnóstico de los riesgos de privacidad de aplicaciones. Ponemos en relación las Políticas de Privacidad y Términos de servicio de las aplicaciones con el cumplimiento de la normativa y los permisos obtenidos, y realmente utilizados en la instalación y uso real de las aplicaciones móviles. Los resultados principales que hemos obtenido son la constitución de los Datasets de privacidad y las métricas de evaluación descritas a lo largo de este trabajo. Nuestra investigación integra el conocimiento de desarrollo propio, los estudios relevantes del sector, además de la aplicación de técnicas de Deep Learning, garantizando un enfoque robusto y basado en las mejores prácticas.

En base a estos resultados identificamos nuevas líneas de desarrollo, como son ampliar el uso de nuestros modelos LLM con funcionalidad interactiva para el usuario, permitiendo evaluar y comprender el funcionamiento de las aplicaciones y los riesgos asociados a su uso. Estas capacidades permiten un enfoque personalizado, según sus expectativas y criterios de seguridad. Nuestros Datasets generados ofrecen una base de conocimiento elaborada con la que los modelos pueden ofrecer una respuesta intuitiva y cercana al lenguaje y necesidades del usuario. Con un enfoque más preventivo, pueden desarrollarse funcionalidades que permitan al usuario establecer niveles de tolerancia, indicadores personalizados y alertas específicas respecto de condiciones de privacidad por parte de las aplicaciones.

Agradecimientos

Los autores desean agradecer el apoyo del Gobierno de España (a través del Instituto Nacional de Ciberseguridad de España, INCIBE), con el Proyecto de Investigación Estratégico “Análisis de aplicaciones móviles desde la perspectiva de la protección de datos: Ciberprotección y Ciberriesgos de la información ciudadana”, así como la Cátedra Internacional “Smart Rural IoT y Entornos Seguros”, en el contexto del Plan de Recuperación, Transformación y Resiliencia financiado por la Unión Europea (NextGenerationEU/PRTR). Los autores también agradecen el soporte del Consorcio del Centro Asociado UNED Ponferrada. INTECCA, así como el grupo de investigación UNED CiberCSI y el grupo de innovación docente UNED CiberGID.

Referencias

[1] I. Pollach, «What’s wrong with online privacy policies?» en, *Communications of the ACM*, vol. 50, n.º 9, págs. 103-108, sep. de 2007, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/1284621.1284627. dirección: https://dl.acm.org/doi/10.1145/1284621.1284627 (visitado 08-09-2024).

[2] T. Linden, R. Khandelwal, H. Harkous y K. Fawaz, *The Privacy Policy Landscape After the GDPR*, arXiv:1809.08396 [cs], jun. de 2019. dirección: http://arxiv.org/abs/1809.08396 (visitado 24-04-2024).

[3] N. N. Kumar y R. Shyamasundar, «Realizing Purpose-Based Privacy Policies Succinctly via Information-Flow Labels,» en *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, Sydney, Australia: IEEE, dic. de 2014, págs. 753-760, ISBN: 978-1-4799-6719-3. DOI: 10.1109/BDCloud.2014.89. dirección: http://ieeexplore.ieee.org/document/7034870/ (visitado 08-09-2024).

[4] Á. Feal, P. Calciati, N. Vallina-Rodriguez, C. Troncoso y A. Gorla, «Angel or Devil? A Privacy Study of Mobile Parental Control Apps,» en, *Proceedings on Privacy Enhancing Technologies*, vol. 2020, n.º 2, págs. 314-335, abr. de 2020, ISSN: 2299-0984. DOI: 10.2478/popets-2020-0029. dirección: https://petsymposium.org/popets/2020/popets-2020-0029.php (visitado 08-09-2024).

[5] Z. Tan y W. Song, «PTPDroid: Detecting Violated User Privacy Disclosures to Third-Parties of Android Apps,» en *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, Melbourne, Australia: IEEE, mayo de 2023, págs. 473-485, ISBN: 978-1-6654-5701-9. DOI: 10.1109/ICSE48619.2023.00050. dirección: https://ieeexplore.ieee.org/document/10172832/ (visitado 08-09-2024).

[6] R. N. Zaeem, R. L. German y K. S. Barber, «PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining,» en, *ACM Transactions on Internet Technology*, vol. 18, n.º 4, págs. 1-18, nov. de 2018, ISSN: 1533-5399, 1557-6051. DOI: 10.1145/3127519. dirección: https://dl.acm.org/doi/10.1145/3127519 (visitado 08-09-2024).

[7] S. Wilson, F. Schaub, A. A. Dara et al., «The Creation and Analysis of a Website Privacy Policy Corpus,» en, en *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, 2016, págs. 1330-1340. DOI: 10.18653/v1/P16-1126. dirección: http://aclweb.org/anthology/P16-1126 (visitado 24-04-2024).

[8] M. S. Rahman, P. Naghavi, B. Kojusner et al., «PermPress: Machine Learning-Based Pipeline to Evaluate Permissions in App Privacy Policies,» *IEEE Access*, vol. 10, págs. 89 248-89 269, 2022, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3199882. dirección: https://ieeexplore.ieee.org/document/9861610/ (visitado 29-04-2024).

[9] R. Mcconkey y O. Olukoya, «Runtime and Design Time Completeness Checking of Dangerous Android App Permissions Against GDPR,» *IEEE Access*, vol. 12, págs. 1-22, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3347194. dirección: https://ieeexplore.ieee.org/document/10373786/ (visitado 29-04-2024).

[10] N. Mousavi Nejad, P. Jabat, R. Nedelchev, S. Scerri y D. Graux, «Establishing a Strong Baseline for Privacy Policy Classification,» en, en *ICT Systems Security and Privacy Protection*, M. Hölbl, K. Rannenberg y T. Welzer, eds., vol. 580, Series Title: IFIP Advances in Information and Communication Technology, Cham: Springer International Publishing, 2020, págs. 370-383, ISBN: 978-3-030-58200-5. DOI: 10.1007/978-3-030-58201-2_25. dirección: https://link.springer.com/10.1007/978-3-030-58201-2_25 (visitado 08-09-2024).

[11] M. Kretschmer, J. Pennekamp y K. Wehrle, «Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web,» en, *ACM Transactions on the Web*, vol. 15, n.º 4, págs. 1-42, nov. de 2021, ISSN: 1559-1131, 1559-114X. DOI: 10.1145/3466722. dirección: https://dl.acm.org/doi/10.1145/3466722 (visitado 08-09-2024).

[12] S. Zimmeck, Z. Wang, L. Zou et al., «Automated Analysis of Privacy Requirements for Mobile Apps,» en, en *Proceedings 2017 Network and Distributed System Security Symposium*, San Diego, CA: Internet Society, 2017, ISBN: 978-1-891562-46-4. DOI: 10.14722/ndss.2017.23034. dirección: https://www.ndss-symposium.org/ndss2017/ndss-2017-programme/automated-analysis-privacy-requirements-mobile-apps/ (visitado 08-09-2024).

[13] L. Yu, X. Luo, J. Chen et al., «PPChecker: Towards Accessing the Trustworthiness of Android Apps’ Privacy Policies,» *IEEE Transactions on Software Engineering*, vol. 47, n.º 2, págs. 221-242, feb. de 2021, ISSN: 0098-5589, 1939-3520, 2326-3881. DOI: 10.1109/TSE.2018.2886875. dirección: https://ieeexplore.ieee.org/document/8576579/ (visitado 08-09-2024).

[14] R. Baalous y R. Poet, «Utilizing Sentence Embedding for Dangerous Permissions Detection in Android Apps' Privacy Policies:» en, *International Journal of Information Security and Privacy*, vol. 15, n.º 1, págs. 173-189, ene. de 2021, ISSN: 1930-1650, 1930-1669. DOI: 10.4018/IJISP.2021010109. dirección: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJISP.2021010109 (visitado 08-09-2024).

[15] R. Slavin, X. Wang, M. B. Hosseini et al., «Toward a framework for detecting privacy policy violations in android application code,» en, en *Proceedings of the 38th International Conference on Software Engineering*, Austin Texas: ACM, mayo de 2016, págs. 25-36, ISBN: 978-1-4503-3900-1. DOI: 10.1145/2884781.2884855. dirección: https://dl.acm.org/doi/10.1145/2884781.2884855 (visitado 08-09-2024).

[16] A. Sunyaev, T. Dehling, P. L. Taylor y K. D. Mandl, «Availability and quality of mobile health app privacy policies,» en, *Journal of the American Medical Informatics Association*, vol. 22, n.º e1, e28-e33, abr. de 2015, ISSN: 1527-974X, 1067-5027. DOI: 10.1136/amiajnl-2013-002605. dirección: https://academic.oup.com/jamia/article/22/e1/e28/700676 (visitado 08-09-2024).

[17] A. Hooda, R. Khandelwal, P. Chalasani, K. Fawaz y S. Jha, *PolicyLR: A Logic Representation For Privacy Policies*, arXiv:2408.14830 [cs], ago. de 2024. dirección: http://arxiv.org/abs/2408.14830 (visitado 08-09-2024).

[18] L. F. Cranor, S. Egelman, S. Sheng, A. M. McDonald y A. Chowdhury, «P3P deployment on websites,» en, *Electronic Commerce Research and Applications*, vol. 7, n.º 3, págs. 274-293, sep. de 2008, ISSN: 15674223. DOI: 10.1016/j.elerap.2008.04.003. dirección: https://linkinghub.elsevier.com/retrieve/pii/S1567422308000185 (visitado 08-09-2024).

[19] S. D. Gupta y T. Hahmann, *OPPO: An Ontology for Describing Fine-Grained Data Practices in Privacy Policies of Online Social Networks*, Version Number: 1, 2023. DOI: 10.48550/ARXIV.2309.15971. dirección: https://arxiv.org/abs/2309.15971 (visitado 08-09-2024).

[20] S. Liu, F. Zhang, B. Zhao, R. Guo, T. Chen y M. Zhang, «APPCorp: a corpus for Android privacy policy document structure analysis,» en, *Frontiers of Computer Science*, vol. 17, n.º 3, pág. 173 320, jun. de 2023, ISSN: 2095-2228, 2095-2236. DOI: 10.1007/s11704-022-1627-2. dirección: https://link.springer.com/10.1007/s11704-022-1627-2 (visitado 13-05-2024).

[21] É. P. F. d. L. (Hamza Harkous, U. o. W.-M. Kassem Fawaz, Rémi Lebret, École Polytechnique Fédérale de Lausanne (EPFL), Rémi Lebret, École Polytechnique Fédérale, Florian Schaub and Kang G. Shin, University of Michigan y É. P. F. d. L. (Karl Aberer, «Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning,» Baltimore, MD, USA, ago. de 2018, ISBN: 978-1-939133-04-5. dirección: https://www.usenix.org/system/files/conference/usenixsecurity18/sec18-harkous.pdf (visitado 24-04-2024).

[22] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini y L. Robaldo, «PrOnto: Privacy Ontology for Legal Reasoning,» en, en *Electronic Government and the Information Systems Perspective*, A. Kő y E. Francesconi, eds., vol. 11032, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, págs. 139-152, ISBN: 978-3-319-98348-6. DOI: 10.1007/978-3-319-98349-3_11. dirección: https://link.springer.com/10.1007/978-3-319-98349-3_11 (visitado 08-09-2024).

[23] A. Oltramari, D. Piraviperumal, F. Schaub et al., «PrivOnto: A semantic framework for the analysis of privacy policies,» *Semantic Web*, vol. 9, n.º 2, M. d'Aquin, S. Kirrane, S. Villata, M. d'Aquin, S. Kirrane y S. Villata, eds., págs. 185-203, ene. de 2018, ISSN: 22104968, 15700844. DOI: 10.3233/SW-170283. dirección: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-170283 (visitado 25-04-2024).

[24] S. Zimmeck, P. Story, D. Smullen et al., «MAPS: Scaling Privacy Compliance Analysis to a Million Apps,» en, *Proceedings on Privacy Enhancing Technologies*, vol. 2019, n.º 3, págs. 66-86, jul. de 2019, ISSN: 2299-0984. DOI: 10.2478/popets-2019-0037. dirección: https://petsymposium.org/popets/2019/popets-2019-0037.php (visitado 06-05-2024).

[25] S. Manandhar, K. Singh y A. Nadkarni, «Towards Automated Regulation Analysis for Effective Privacy Compliance,» en, en *Proceedings 2024 Network and Distributed System Security Symposium*, San Diego, CA, USA: Internet Society, 2024, ISBN: 978-1-891562-93-8. DOI: 10.14722/ndss.2024.24650. dirección: https://www.ndss-symposium.org/wp-content/uploads/2024-650-paper.pdf (visitado 06-02-2025).

[26] A. Xiang, W. Pei y C. Yue, «PolicyChecker: Analyzing the GDPR Completeness of Mobile Apps' Privacy Policies,» en, en *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, Copenhagen Denmark: ACM, nov. de 2023, págs. 3373-3387, ISBN: 979-8-4007-0050-7. DOI: 10.1145/3576915.3623067. dirección: https://dl.acm.org/doi/10.1145/3576915.3623067 (visitado 06-02-2025).

[27] A. Ravichander, A. W. Black, S. Wilson, T. Norton y N. Sadeh, «Question Answering for Privacy Policies: Combining Computational and Legal Perspectives,» 2019. DOI: 10.48550/ARXIV.1911.00841. dirección: https://arxiv.org/abs/1911.00841 (visitado 02-05-2024).

[28] W. U. Ahmad, J. Chi, Y. Tian y K.-W. Chang, *PolicyQA: A Reading Comprehension Dataset for Privacy Policies*, Version Number: 1, 2020. DOI: 10.48550/ARXIV.2010.02557. dirección: https://arxiv.org/abs/2010.02557 (visitado 08-09-2024).

[29] H. Habib, S. Pearman, J. Wang et al., «It's a scavenger hunt: Usability of Websites' Opt-Out and Data Deletion Choices,» en, en *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA: ACM, abr. de 2020, págs. 1-12, ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376511. dirección: https://dl.acm.org/doi/10.1145/3313831.3376511 (visitado 08-09-2024).

[30] C. Jensen y C. Potts, «Privacy policies as decision-making tools: an evaluation of online privacy notices,» en, en *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna Austria: ACM, abr. de 2004, págs. 471-478, ISBN: 978-1-58113-702-6. DOI: 10.1145/985692.985752. dirección: https://dl.acm.org/doi/10.1145/985692.985752 (visitado 08-09-2024).

[31] AppBrain, *Most Popular Social Apps - AppBrain*, Accessed: 03-Mar-2025, 2024. dirección: https://www.appbrain.com/apps/popular/social.

[32] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun e Y. Zhang, «A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly,» en, *High-Confidence Computing*, vol. 4, n.º 2, pág. 100 211, jun. de 2024, ISSN: 26672952. DOI: 10.1016/j.hcc.2024.100211. dirección: https://linkinghub.elsevier.com/retrieve/pii/S266729522400014X (visitado 07-02-2025).

[33] D. Blanco-Aza, A. Robles-Gómez, R. Pastor-Vargas, L. Tobarra, P. Vidal-Balboa y M. Méndez-Suárez, «Privacy Analysis in Mobile Apps and Social Networks Using AI Techniques,» en *2024 4th Intelligent Cybersecurity Conference (ICSC)*, Valencia, Spain: IEEE, sep. de 2024, págs. 188-195, ISBN: 9798350354775. DOI: 10.1109/ICSC63108.2024.10895037. dirección: https://ieeexplore.ieee.org/document/10895037/ (visitado 04-03-2025).

[34] A. Adhikari, S. Das y R. Dewri, «Evolution of Composition, Readability, and Structure of Privacy Policies over Two Decades,» *Proceedings on Privacy Enhancing Technologies*, vol. 2023, n.º 3, págs. 138-153, jul. de 2023, ISSN: 2299-0984. DOI: 10.56553/popets-2023-0074. dirección: https://petsymposium.org/popets/2023/popets-2023-0074.php (visitado 08-09-2024).

El espejismo de la transparencia: Discrepancias en las declaraciones de privacidad en apps móviles

Pablo A. Criado-Lozano
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid
Valladolid, España
Universidad Europea Miguel de Cervantes
Valladolid, España
pacriado@uemc.es

M. Mercedes Martínez-González
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid
Valladolid, España
mercedes@infor.uva.es

Alejandro Pérez-Fuente
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid
Valladolid, España
alejandro.perez.fuente@uva.es

Amador Aparicio
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid
Valladolid, España
amador@infor.uva.es

***Resumen*—La transparencia en el uso de datos personales en aplicaciones móviles es fundamental para garantizar la confianza del usuario. Sin embargo, la coherencia entre las distintas declaraciones de privacidad asociadas a una misma aplicación —la política de privacidad, la sección de Seguridad de los Datos en Google Play y los permisos solicitados en el archivo AndroidManifest.xml— no siempre está asegurada, lo que puede comprometer la fiabilidad de la información proporcionada a los usuarios. Este estudio propone un método sistemático para evaluar la coherencia entre estas tres fuentes de información. Como prueba de concepto, se ha aplicado el método a Samsung Health. Se han identificado discrepancias entre las fuentes de información. Los resultados muestran que es necesario realizar comprobaciones que garanticen la fiabilidad y la coherencia de la información ofrecida a los usuarios. Esto contribuiría a una mayor calidad en la comunicación de las prácticas de tratamiento de datos y, en última instancia, a una mejor transparencia.**

***Index Terms*—Privacidad, transparencia, aplicaciones móviles, seguridad de los datos, permisos, Android.**

Tipo de contribución: *Investigación original*

I. Introducción

El crecimiento del uso de aplicaciones móviles ha generado preocupaciones sobre la privacidad y seguridad de los datos personales, ya que muchas recopilan y procesan información sensible, como datos de salud, ubicación, contactos y hábitos de uso, con el riesgo de exposición y uso indebido. En respuesta a estas preocupaciones, marcos regulatorios como el *Reglamento General de Protección de Datos* (RGPD)[1] en Europa han establecido normativas para la recopilación, almacenamiento y tratamiento de datos personales. No obstante, la implementación efectiva de estos principios de transparencia y control por parte de los usuarios sigue siendo un desafío, especialmente en los ecosistemas móviles.

Uno de los principales desafíos en la evaluación de la privacidad de las aplicaciones móviles radica en la discrepancia entre la información que una aplicación declara sobre la recopilación y el uso de datos personales y los permisos que solicita en el dispositivo. Esta falta de correspondencia incumple con el deber de informar claramente a los usuarios sobre el tratamiento efectivo de sus datos personales, afectando negativamente a la transparencia exigida por los organismos supervisores [2], [3]. Como consecuencia, los usuarios pueden experimentar dificultades significativas para comprender con precisión qué datos están siendo recopilados y con qué propósito, lo que limita su capacidad para tomar decisiones plenamente informadas sobre la protección de su privacidad.

Para abordar esta problemática, Google Play ha implementado la *Sección de Seguridad de los Datos*, un apartado en el que, desde julio de 2022, los desarrolladores deben especificar qué tipos de información recopilan, cómo se utilizan y si los datos se comparten con terceros[4]. Sin embargo, a pesar de estos esfuerzos, el problema de coherencia y fiabilidad de la información sigue presente.

Los usuarios reciben información desde tres fuentes distintas: la política de privacidad, la sección de Seguridad de los Datos en Google Play y los permisos solicitados en el archivo AndroidManifest.xml. Ninguna de estas fuentes garantiza por sí sola la veracidad y precisión de la información que presentan, lo que puede generar confusión y afectar la transparencia. En consecuencia, es necesario avanzar hacia una mejor coherencia y calidad de la información que reciben los usuarios, asegurando que estas fuentes sean consistentes entre sí y reflejen con mayor precisión el uso real de los datos personales.

Se presenta un metodo de análisis para identificar y evaluar discrepancias entre tres fuentes clave de información en las aplicaciones móviles:

- **La política de privacidad de la aplicación**, donde los desarrolladores declaran el uso de datos personales.
- **La sección de seguridad de los datos en Google Play**, que proporciona información sobre los datos recopilados.
- **Los permisos solicitados por la aplicación en el archivo** `AndroidManifest.xml`, que determinan los recursos a los que puede acceder en el dispositivo.

Definimos una discrepancia como cualquier inconsistencia entre la información que una aplicación móvil declara en su política de privacidad, lo reportado en la sección de seguridad de los datos en Google Play y los permisos solicitados en el archivo `AndroidManifest.xml`.

Hasta donde conocemos, no existe un método previo que verifique la transparencia de una aplicación móvil contrastando estas tres fuentes simultáneamente. Esto resalta la relevancia de nuestra propuesta como una contribución novedosa para evaluar la coherencia entre lo que se declara y lo que se solicita en términos de datos personales.

Este estudio se centra en la evaluación de aplicaciones que acceden a datos de salud, dado que esta información es particularmente sensible en términos de privacidad. La aparición de *Health Connect* en Android, aunque permite una gestión más granular de estos datos, no elimina la necesidad de garantizar la coherencia y fiabilidad en las declaraciones de privacidad. El análisis de discrepancias en este tipo de aplicaciones es relevante.

Hemos realizado una prueba de concepto con la aplicación *Samsung Health* en su versión 6.29.2.001.

A partir de esta evaluación, se busca responder a la siguiente pregunta. ¿Existen discrepancias significativas entre la información declarada por los desarrolladores en sus políticas de privacidad y en Google Play respecto a los permisos solicitados por la aplicación?

La estructura del artículo es la siguiente: en la sección II se presenta el estado del arte; en la sección III se revisa el trabajo relacionado; en la sección IV se describe la propuesta metodológica; en la sección V se presentan los experimentos realizados; en la sección VI se analizan los resultados obtenidos; en la sección VII se exponen las conclusiones y, por último, en la sección VIII se indican las líneas de trabajo futuro

II. Estado del arte

El modelo de permisos en Android ha evolucionado con el tiempo, pasando de un sistema en el que los permisos eran otorgados en el momento de la instalación a un modelo basado en permisos dinámicos. A pesar de estos avances, el modelo sigue presentando desafíos para la evaluación de privacidad, especialmente en aplicaciones que recopilan grandes volúmenes de datos personales.

En Android, un permiso generalmente otorga acceso a un recurso sin especificar los datos concretos dentro de dicho recurso. Por ejemplo:

- **Permisos de almacenamiento** (`READ_EXTERNAL_STORAGE`, `WRITE_EXTERNAL_STORAGE`) permiten acceder a cualquier archivo en el almacenamiento del usuario, sin distinción entre fotos personales, documentos o datos generados por otras aplicaciones.
- **Permisos de ubicación** (`ACCESS_FINE_LOCATION`, `ACCESS_COARSE_LOCATION`) otorgan acceso a la geolocalización del usuario, pero no distinguen entre una ubicación específica usada para navegación y datos históricos de movimiento.
- **Permisos de sensores** (`BODY_SENSORS`) permiten acceder a datos biométricos, pero sin especificar si la aplicación leerá la frecuencia cardíaca, la saturación de oxígeno o la presión arterial.

El mero hecho de que una aplicación solicite un permiso no implica que acceda a todos los datos posibles dentro de ese permiso[5]. Esta limitación ha sido un obstáculo en la evaluación de la privacidad de las aplicaciones móviles, hasta la introducción de la API *Health Connect* de Google.

II-1. Health Connect: Un modelo de permisos más granular: Las aplicaciones de salud han experimentado un crecimiento significativo en los últimos años, proporcionando herramientas avanzadas para el monitoreo y gestión del bienestar de los usuarios. Estudios recientes han demostrado que muchas de estas aplicaciones manejan información sensible, incluyendo datos biométricos, frecuencia cardíaca, presión arterial y actividad física [6], [7]. En entornos de *Internet de Todo* (IoE), donde los dispositivos y sensores están interconectados, la protección de estos datos se vuelve aún más crítica.

En respuesta a estas preocupaciones, Google introdujo *Health Connect*, una API que permite la gestión centralizada de datos de salud en dispositivos Android. A diferencia del modelo tradicional de permisos en Android, en el cual un permiso otorgaba acceso a un recurso completo sin especificar qué datos concretos serían utilizados, Health Connect permite que las aplicaciones soliciten **permisos granulares**, limitando el acceso a tipos específicos de datos personales. Por ejemplo:

- `READ_HEART_RATE`: Permite acceso únicamente a la frecuencia cardíaca.
- `READ_BLOOD_GLUCOSE`: Restringe el acceso a los niveles de glucosa en sangre sin habilitar otros datos biométricos.
- `READ_SLEEP`: Controla exclusivamente el acceso a información sobre patrones de sueño.

Este nivel de control mejora la transparencia en el uso de datos sensibles porque permite que los usuarios otorguen permisos de manera más específica, en lugar de conceder acceso amplio a categorías completas de datos. Al poder seleccionar exactamente qué tipo de información pueden recopilar las aplicaciones (por ejemplo, frecuencia cardíaca sin necesidad de compartir otros datos biométricos), se reduce la ambigüedad sobre el alcance del acceso a los datos personales. Esto facilita a los usuarios la gestión de su privacidad de manera más efectiva, al brindarles mayor visibilidad y control sobre la información que comparten.

Google Play ha actualizado sus políticas sobre acceso a información sensible, restringiendo desde mayo de 2025 el uso de permisos generales para fotos y videos, y estableciendo directrices más claras sobre datos médicos mediante la sección *Health Content and Services* [8], [9], sin embargo, estas medidas aún dependen de la autorregulación de los desarrolladores, lo que cuestiona su efectividad en términos de transparencia y gestión coherente de los datos sensibles.

II-A. Ejemplo de la declaración en las diferentes fuentes: Samsung Health y la frecuencia cardíaca

Para ilustrar las diferencias en la declaración de un mismo dato en las distintas fuentes de información, se ha tomado como referencia la frecuencia cardíaca en la aplicación **Samsung Health**. En la Figura 1, se muestra cómo este dato

es declarado en su **Política de Privacidad**. En la Figura 2, se observa cómo la misma información se presenta en la sección de **Seguridad de los Datos en Google Play**, y en la Figura 3, se visualiza el permiso correspondiente en el **AndroidManifest.xml**.

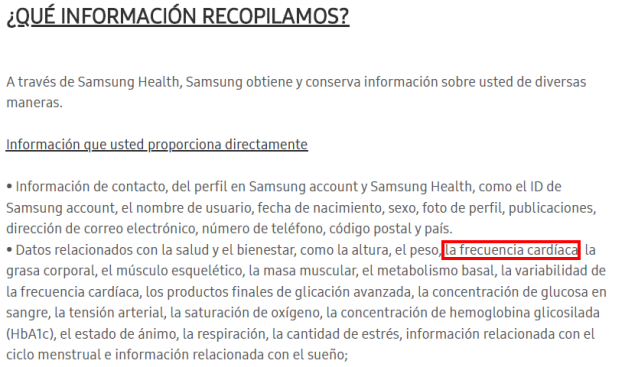


Figura 1. Declaración de la frecuencia cardíaca en la Política de Privacidad de Samsung Health.



Figura 2. Declaración de la frecuencia cardíaca en la sección de Seguridad de los Datos en Google Play.

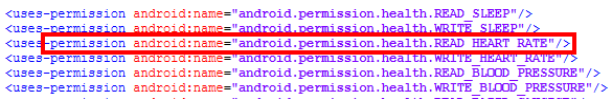


Figura 3. Permiso de lectura de la frecuencia cardíaca en el archivo AndroidManifest.xml.

Existen diferencias en la forma en que se presenta la recopilación de la frecuencia cardíaca. Mientras que en la Política de Privacidad se especifica que se accede a la frecuencia cardíaca, en la sección de Seguridad de los Datos en Google Play se presenta de forma más resumida y generalizada dentro de la categoría de "información de salud". Por otro lado, en el AndroidManifest.xml, el permiso `android.permission.health.READ_HEART_RATE` especifica directamente qué tipo de dato será accedido.

Es importante destacar que las **políticas de privacidad** y la seccion **seguridad de datos de google play**, están dirigidas a los usuarios, mientras que el archivo `AndroidManifest.xml` es para consumo del **sistema operativo**. Las primeras, presentan la información en un lenguaje accesible y comprensible para los usuarios.

Por otro lado, el `AndroidManifest.xml` no está destinado a los usuarios finales, sino al sistema operativo Android. Este archivo define de manera **técnica y estructurada** los permisos requeridos por la aplicación para acceder a distintos recursos del dispositivo. A diferencia de las políticas de privacidad, aquí no se explica el propósito del acceso a los datos, sino que simplemente se declara la necesidad de un permiso específico para su funcionamiento.

Esta distinción es fundamental para comprender el problema de las discrepancias: **la falta de alineación entre las declaraciones diseñadas para los usuarios y la realidad técnica de los permisos** puede generar confusión. En consecuencia, la metodología propuesta busca cerrar esta brecha mediante un *análisis sistemático* que permita detectar inconsistencias entre ambas perspectivas.

III. Trabajo Relacionado

La privacidad en aplicaciones móviles ha sido ampliamente estudiada, especialmente en relación con el tratamiento de datos personales y el acceso a información sensible, como los datos de salud. En este estudio, nos centraremos en dos puntos de vista para estudiar la privacidad en las aplicaciones móviles:

- Discrepancia en el uso de datos y la privacidad
- Riesgos de privacidad en aplicaciones de salud

III-A. Discrepancia en el uso de datos y la privacidad

La evaluación de la privacidad en aplicaciones móviles ha sido abordada desde distintos enfoques. Algunos estudios han analizado cómo las aplicaciones acceden y comparten información sin el conocimiento explícito del usuario. Por ejemplo, **Enck et al.**[10] introdujeron TaintDroid, un sistema de monitoreo de flujo de datos en Android que permite rastrear el acceso y la transmisión de datos personales a servidores externos, revelando casos en los que la información del usuario era compartida sin su consentimiento explícito.

Otros enfoques han centrado su análisis en la coherencia entre las declaraciones de privacidad y los permisos solicitados por las aplicaciones. **Slavin et al.**[11] desarrollaron un método automatizado para comparar los permisos declarados en el archivo `AndroidManifest.xml` con la información proporcionada en la política de privacidad, identificando discrepancias significativas donde las aplicaciones recopilaban más datos de los indicados. De manera similar, **Zimmeck et al.**[12] evidenciaron inconsistencias en la forma en que las aplicaciones comunican su uso de datos a los usuarios, lo que pone en entredicho la transparencia de estas declaraciones.

Yu et al. [13] propusieron un sistema basado en análisis de código y procesamiento de lenguaje natural para detectar discrepancias en políticas de privacidad. Sus hallazgos revelaron que un porcentaje significativo de aplicaciones no declaraba correctamente los datos que recopilaban, lo que revela una falta de transparencia en la industria de aplicaciones móviles.

No obstante, una limitación común en estos estudios es la falta de granularidad en la evaluación de los permisos y los datos específicos a los que acceden las aplicaciones.

III-B. Riesgos de privacidad en aplicaciones de salud

A pesar de las mejoras en la gestión de permisos, las aplicaciones de salud siguen representando un desafío en términos de privacidad y seguridad. La recopilación de datos biométricos, información médica y hábitos de salud plantea

riesgos significativos si no se garantiza una gestión adecuada. Estudios previos han analizado estos riesgos desde distintas perspectivas, incluyendo el análisis legal de políticas de privacidad, la revisión técnica de permisos y la evaluación de la transferencia de datos personales [14], [15].

Antes de la introducción de *Health Connect*, la evaluación de la privacidad en aplicaciones de salud se realizaba sobre un modelo de permisos más amplio y menos granular. En este contexto, **Tangari et al.** [16] realizaron un análisis a gran escala de aplicaciones de salud en Google Play, encontrando que muchas de ellas transmitían datos sensibles sin cifrado adecuado e incluso algunas contenían malware. De manera similar, **Papageorgiou et al.**[7] analizaron la seguridad y privacidad de aplicaciones móviles de salud y detectaron múltiples casos de uso indebido de datos personales, incluyendo la falta de transparencia en la comunicación con los usuarios. Estos hallazgos resaltan la necesidad de evaluar con mayor rigurosidad la seguridad de este tipo de aplicaciones.

Tras la introducción de *Health Connect*, algunas investigaciones han comenzado a estudiar el impacto de esta API en la privacidad de los usuarios. **Martínez-González et al.** [6], en el marco del ecosistema App-PI, han propuesto un enfoque basado en metadatos para evaluar la transparencia en aplicaciones de salud. Sus hallazgos sugieren que, aunque los permisos de *Health Connect* proporcionan una mayor granularidad y control, todavía existe una brecha entre la información que los usuarios reciben y el acceso real que las aplicaciones tienen a sus datos.

Algunos estudios señalan limitaciones importantes en la transparencia y precisión con la que las aplicaciones móviles de salud informan a los usuarios sobre los datos recopilados y su finalidad. **Benjumea et al.** [15] llevaron a cabo una revisión sobre cómo se evalúa la privacidad en aplicaciones móviles de salud, indicando que muchas aplicaciones continúan sin informar claramente sobre el tratamiento y propósito de los datos personales que recopilan. Asimismo, **Yu et al.** [13] evidenciaron mediante análisis automatizados la falta de correspondencia entre las declaraciones contenidas en las políticas de privacidad y los datos que recopilan las aplicaciones, afectando directamente la confianza del usuario y su capacidad para gestionar adecuadamente su información personal.

IV. Hacia una Mayor Transparencia: Método para Evaluar Discrepancias en las Declaraciones

Para conseguir un *método sistemático y reproducible* que permita detectar las discrepancias entre las declaraciones de privacidad de las aplicaciones móviles y los permisos solicitados, se propone una metodología estructurada en tres fases.

A continuación, se detallan las fuentes de datos utilizadas en el análisis y las fases que conforman la metodología propuesta.

IV-A. Fuentes de datos

En esta metodología se analizan tres fuentes fundamentales de información:

- La **política de privacidad** de la aplicación, donde los desarrolladores están obligados por normativa [1] a declarar explícitamente qué datos personales recopilan, almacenan y comparten, así como el propósito de dicho tratamiento.
- La **sección de Seguridad de los Datos en Google Play**, que presenta una versión simplificada de los datos personales recopilados, orientada al usuario final.
- Los **permisos solicitados en el archivo `AndroidManifest.xml`**, que determinan técnicamente los recursos específicos del dispositivo y los datos personales a los que puede acceder la aplicación, especialmente mediante la API *Health Connect*.

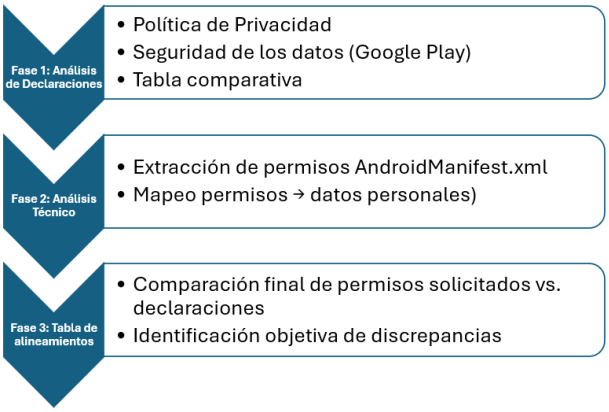


Figura 4. Diagrama de la metodología.

IV-B. Fases de análisis

La metodología propuesta se estructura en tres fases, tal y como se muestra en la Figura 4.

Fase 1: Análisis de declaraciones (política de privacidad y Seguridad de los Datos en Google Play)

En primer lugar, se realiza un análisis de los documentos proporcionados por el desarrollador:

1. Se extrae información específica de la *política de privacidad*, identificando todos los datos personales mencionados, poniendo énfasis en datos de salud.
2. Se realiza una extracción de la información proporcionada en la *sección de Seguridad de los Datos de Google Play*, identificando los tipos de datos declarados como recopilados o compartidos por la aplicación.
3. A partir de estas extracciones, se construye una *tabla comparativa* para facilitar la detección de las discrepancias existentes entre ambas fuentes documentales. Los resultados se muestran en la Tabla I

Fase 2: Análisis técnico (permisos en AndroidManifest.xml)

En segundo lugar, se lleva a cabo un análisis mediante inspección del archivo `AndroidManifest.xml`, obteniendo una perspectiva clara de los permisos solicitados por la aplicación:

1. Se extrae la lista completa de todos los permisos solicitados por la aplicación, tal y como se puede ver en la Tabla II.
2. Se ha realizado un **mapeo automatizado** de los permisos de *Health Connect* hacia los datos personales concretos que pueden ser accedidos, utilizando como

referencia la documentación oficial [17] de Google. Para el resto de datos personales, como ya se indicó en la Sección II, *Android no establece una relación directa entre los permisos y los datos a los que pueden acceder las aplicaciones*. Debido a esta limitación, se ha realizado una **extracción supervisada**, apoyada en la documentación oficial de Android, con el fin de garantizar una asignación precisa y justificada. Este mapeo puede verse en la Tabla III.

Fase 3: Generación de tabla de alineamientos

Finalmente, se genera una *tabla de alineamientos*, similar a la mostrada en la Tabla VI, en la que se contrastan los permisos obtenidos del archivo `AndroidManifest.xml` con los datos declarados en la política de privacidad y en la sección de Seguridad de los Datos en Google Play. Esta tabla permite la identificación discrepancias entre las declaraciones documentales y los accesos técnicos solicitados por la aplicación.

Para la construcción de la tabla, se han definido los siguientes atributos o columnas:

- **Permiso**: Nombre del permiso solicitado en el archivo `AndroidManifest.xml`.
- **Dato personal asociado**: Información personal que el permiso podría permitir acceder.
- **Presencia en la política de privacidad**: Indica si el dato personal asociado está declarado en la política de privacidad de la aplicación.
- **Presencia en la sección de Seguridad de los Datos en Google Play**: Indica si el dato personal está reportado en Google Play como recopilado por la aplicación.
- **Coherencia entre política de privacidad y Google Play**: Evalúa si la información declarada en ambas fuentes es consistente.
- **Coherencia entre Google Play y los permisos**: Evalúa si los datos reportados en Google Play reflejan los permisos solicitados en `AndroidManifest.xml`.
- **Coherencia entre la política de privacidad y los permisos**: Indica si los permisos solicitados coinciden con los datos que la política de privacidad declara recopilar.

Cada fila de la tabla representa un **dato personal**, es decir, los valores de cada fila corresponden a una categoría específica de información que puede ser accedida por la aplicación. La decisión de estructurar la tabla de esta manera responde a la necesidad de evaluar la transparencia desde la perspectiva del usuario, facilitando la identificación de discrepancias en la manera en que se comunican los datos personales en cada una de las fuentes analizadas.

V. Samsung Health bajo la lupa: evaluando su transparencia

Como se ha indicado en la sección I, se ha seleccionado la aplicación *Samsung Health* en su versión *6.29.2.001*.

Para garantizar la reproducibilidad del experimento, se han puesto a disposición el APK de la aplicación, el archivo `AndroidManifest.xml`, así como la política de privacidad y la sección de seguridad de los datos de Google Play [18]. Adicionalmente se ha puesto tambien a disposicion la tabla VI, ya que en este artículo se muestra unicamente una parte de la misma.

Fase 1: Análisis de declaraciones

Se ha analizado la política de privacidad de *Samsung Health* [19], identificando los datos personales declarados y comparándolos con la información obtenida de la sección *Seguridad de los Datos* de Google Play, extraída mediante el ecosistema *App-PI* [6]. Los resultados comparativos se presentan en la Tabla I.

Se observa que la política de privacidad proporciona mayor detalle sobre los datos personales recopilados que la sección *Seguridad de los Datos* de Google Play, la cual agrupa los datos en categorías generales.

Tabla I
Datos personales en la política de privacidad y en la seguridad de datos de Google Play de la aplicación Samsung Health

Dato personal	Política	Google Play	Dato personal	Política	Google Play
ID Samsung	Sí	Sí (IDs usr)	Actividad	Sí	Sí (act. física)
Usuario	Sí	Sí (IDs usr)	Velocidad	Sí	Sí (act. física)
Email	Sí	**No**	Pasos	Sí	Sí (act. física)
Teléfono	Sí	Sí	Distancia	Sí	Sí (act. física)
ID publicidad móvil	Sí	Sí (IDs disp.)	Agua	Sí	Sí (act. física)
Nacimiento	Sí	**No**	Foto perfil	Sí	Sí (Fotos)
Sexo	Sí	**No**	Publicaciones	Sí	**No**
Código postal	Sí	**No**	Mensajes Retos	Sí	**No**
País	Sí	**No**	Amigos Retos	Sí	**No**
Altura	Sí	Sí (salud)	Contactos	Sí	Sí (Contactos)
Peso	Sí	Sí (salud)	Ubicación	Sí	Sí (precisa)
Frec. cardíaca	Sí	Sí (salud)	MAC	Sí	**No**
Grasa corporal	Sí	Sí (salud)	Disp. vinculados	Sí	Sí (IDs disp.)
Músculo esq.	Sí	Sí (salud)	Interacc. app	Sí	Sí (interacc.)
Masa muscular	Sí	Sí (salud)	Bluetooth/Wi-Fi	Sí	**No**
Metab. basal	Sí	Sí (salud)	Núm. serie disp.	Sí	**No**
Glucosa sangre	Sí	Sí (salud)	Reg. fallos	Sí	Sí (fallos)
Sat. oxígeno	Sí	Sí (salud)	Diagnósticos	Sí	Sí (diagn.)
Ciclo menstrual	Sí	Sí (salud)	Voz/sonido	**No**	Sí (Audio)
Var. frec. cardíaca	Sí	Sí (salud)	Archivos/docs.	**No**	Sí (Arch./doc.)
Tensión arterial	Sí	Sí (salud)	Estado ánimo	Sí	Sí (salud)
Sueño	Sí	Sí (salud)			

Fase 2: Análisis técnico

Tras extraer la lista completa de permisos solicitados por *Samsung Health* del archivo `AndroidManifest.xml`, se identificaron **76 permisos**, siendo **34 relacionados específicamente con datos de salud**. La Tabla II muestra los permisos identificados.

Tabla II
Permisos Android solicitados por Samsung Health versión 6.29.2.001

N°	Permiso	N°	Permiso
1	VIBRATE	39	AUTHENTICATE_ACCOUNTS
2	WRITE_SETTINGS	40	READ_STEPS
3	POST_NOTIFICATIONS	41	WRITE_STEPS
4	WRITE_EXTERNAL_STORAGE	42	READ_TOTAL_CALORIES_BURNED
5	READ_EXTERNAL_STORAGE	43	WRITE_TOTAL_CALORIES_BURNED
6	READ_MEDIA_IMAGES	44	READ_DISTANCE
7	READ_MEDIA_VISUAL_USER_SELECTED	45	WRITE_DISTANCE
8	ACCESS_FINE_LOCATION	46	READ_VO2_MAX
9	ACCESS_COARSE_LOCATION	47	WRITE_VO2_MAX
10	WAKE_LOCK	48	READ_POWER
11	INTERNET	49	WRITE_POWER
12	FOREGROUND_SERVICE	50	READ_SPEED
13	FOREGROUND_SERVICE_MICROPHONE	51	WRITE_SPEED
14	ACCESS_NETWORK_STATE	52	READ_EXERCISE
15	BODY_SENSORS	53	WRITE_EXERCISE
16	CAMERA	54	READ_NUTRITION
17	READ_CONTACTS	55	WRITE_NUTRITION
18	READ_PHONE_NUMBERS	56	READ_SLEEP
19	READ_PHONE_STATE	57	WRITE_SLEEP
20	RECEIVE_BOOT_COMPLETED	58	READ_HEART_RATE
21	BLUETOOTH	59	WRITE_HEART_RATE
22	BLUETOOTH_ADMIN	60	READ_BLOOD_PRESSURE
23	BLUETOOTH_SCAN	61	WRITE_BLOOD_PRESSURE
24	BLUETOOTH_CONNECT	62	READ_BLOOD_GLUCOSE
25	NFC	63	WRITE_BLOOD_GLUCOSE
26	GET_ACCOUNTS	64	READ_OXYGEN_SATURATION
27	ACTIVITY_RECOGNITION	65	WRITE_OXYGEN_SATURATION
28	HIGH_SAMPLING_RATE_SENSORS	66	READ_WEIGHT
29	SCHEDULE_EXACT_ALARM	67	WRITE_WEIGHT
30	FOREGROUND_SERVICE_MEDIA_PLAYBACK	68	READ_HEIGHT
31	FOREGROUND_SERVICE_SPECIAL_USE	69	WRITE_HEIGHT
32	FOREGROUND_SERVICE_LOCATION	70	READ_BODY_FAT
33	FOREGROUND_SERVICE_HEALTH	71	WRITE_BODY_FAT
34	FOREGROUND_SERVICE_CONNECTED_DEVICE	72	READ_BASAL_METABOLIC_RATE
35	SYSTEM_ALERT_WINDOW	73	WRITE_BASAL_METABOLIC_RATE
36	RECORD_AUDIO	74	READ_SYNC_STATS
37	QUERY_ALL_PACKAGES	75	READ_SYNC_SETTINGS
38	ACCESS_WIFI_STATE	76	WRITE_SYNC_SETTINGS

Posteriormente, se realizó la asociación entre los permisos identificados y los datos personales específicos a los que otorgan acceso. Esta asociación se realizó tomando como referencia la documentación oficial de Android y *Health Connect* [17], dando lugar a la Tabla III.

Cabe destacar que algunos permisos permiten a la aplicación no solo acceder, sino también modificar ciertos datos personales sensibles, como la frecuencia cardíaca y los niveles de glucosa en sangre.

Tabla III
RELACIÓN ENTRE PERMISOS DE SAMSUNG HEALTH Y TIPOS DE DATOS PERSONALES

Permiso	Dato personal	Permiso	Dato personal
WRITE_EXTERNAL_STORAGE	Archivos personales	WRITE_POWER	Potencia de ejercicio
READ_EXTERNAL_STORAGE	Archivos personales	READ_SPEED	Velocidad
READ_MEDIA_IMAGES	Imágenes personales	WRITE_SPEED	Velocidad
READ_MEDIA_VISUAL_USER_SELECTED	Imágenes personales	READ_EXERCISE	Datos de ejercicio
ACCESS_FINE_LOCATION	Ubicación precisa	WRITE_EXERCISE	Datos de ejercicio
ACCESS_COARSE_LOCATION	Ubicación aproximada	READ_NUTRITION	Datos de nutrición
FOREGROUND_SERVICE_MICROPHONE	Voz del usuario	WRITE_NUTRITION	Datos de nutrición
BODY_SENSORS	Datos biométricos	READ_SLEEP	Datos de sueño
CAMERA	Imágenes personales	WRITE_SLEEP	Datos de sueño
READ_CONTACTS	Contactos	READ_HEART_RATE	Ritmo cardíaco
READ_PHONE_NUMBERS	Número teléfono	WRITE_HEART_RATE	Ritmo cardíaco
READ_PHONE_STATE	Estado dispositivo	READ_BLOOD_PRESSURE	Presión arterial
GET_ACCOUNTS	Cuentas usuario	WRITE_BLOOD_PRESSURE	Presión arterial
ACTIVITY_RECOGNITION	Actividad física	READ_BLOOD_GLUCOSE	Glucosa sangre
HIGH_SAMPLING_RATE_SENSORS	Datos biométricos	WRITE_BLOOD_GLUCOSE	Glucosa sangre
RECORD_AUDIO	Voz del usuario	READ_OXYGEN_SATURATION	Oxígeno sangre
READ_STEPS	Actividad física	WRITE_OXYGEN_SATURATION	Oxígeno sangre
WRITE_STEPS	Actividad física	READ_WEIGHT	Peso corporal
READ_TOTAL_CALORIES_BURNED	Calorías quemadas	WRITE_WEIGHT	Peso corporal
WRITE_TOTAL_CALORIES_BURNED	Calorías quemadas	READ_HEIGHT	Altura corporal
READ_DISTANCE	Distancia recorrida	WRITE_HEIGHT	Altura corporal
WRITE_DISTANCE	Distancia recorrida	READ_BODY_FAT	Grasa corporal
READ_VO2_MAX	VO2 máximo	WRITE_BODY_FAT	Grasa corporal
WRITE_VO2_MAX	VO2 máximo	READ_BASAL_METABOLIC_RATE	Metab. basal
READ_POWER	Potencia ejercicio	WRITE_BASAL_METABOLIC_RATE	Metab. basal

Fase 3: Generación de tabla de alineamientos

Finalmente, la Tabla VI presenta un análisis comparativo que evidencia las discrepancias y similitudes entre los permisos solicitados por *Samsung Health*, los datos personales declarados en la política de privacidad y aquellos mencionados en la sección de *Seguridad de los Datos* de Google Play.

Destacan casos en los que datos como la **foto de perfil** o el **correo electrónico** aparecen declarados en la política de privacidad, pero carecen de permisos específicos asociados en el archivo `AndroidManifest.xml`. Esto sugiere que dichos datos podrían recopilarse mediante otros mecanismos, como la interacción directa con la aplicación.

VI. RESULTADOS

Los resultados obtenidos tras aplicar la metodología propuesta a Samsung Health muestran una serie de discrepancias entre la información declarada en la *política de privacidad*, la sección de *seguridad de los datos en Google Play* y los permisos solicitados en el archivo `AndroidManifest.xml`. Estas diferencias pueden deberse a varios factores, como la naturaleza de los permisos en Android, el nivel de detalle proporcionado en la política de privacidad y la forma en que Google Play presenta la seguridad de los datos a los usuarios.

VI-A. Discrepancias derivadas de la naturaleza técnica de los permisos en Android

Un aspecto clave identificado durante el análisis es que no todos los permisos en Android implican necesariamente un acceso directo o explícito a datos personales. Muchos de estos permisos simplemente permiten que la aplicación utilice ciertos recursos del dispositivo, lo que no siempre conlleva la recopilación o el tratamiento de información personal.

Por ejemplo, los siguientes permisos no necesariamente indican la recopilación o tratamiento de datos personales:

- `VIBRATE` o `WAKE_LOCK` , que únicamente permiten funciones técnicas como la vibración o la gestión de energía del dispositivo, respectivamente.
- `WRITE_EXTERNAL_STORAGE` y `READ_EXTERNAL_STORAGE` , que posibilitan el acceso genérico al almacenamiento externo, pero no necesariamente implican recopilación de información personal, dependiendo del uso específico dado por la aplicación.

Esta situación explica parcialmente las discrepancias identificadas, resaltando que la mera presencia de un permiso en el archivo `AndroidManifest.xml` no implica automáticamente un uso o recopilación indebida de datos personales. Sin embargo, para garantizar una transparencia efectiva, es fundamental que tanto la política de privacidad como la sección Seguridad de los datos en Google Play aclaren explícitamente cuál es el propósito específico de cada permiso solicitado, incluso aquellos aparentemente técnicos.

VI-B. Diferencias entre la Política de Privacidad y la seguridad de los datos en Google Play

Uno de los hallazgos más relevantes es la diferencia en el nivel de detalle entre la política de privacidad y la seguridad de los datos en Google Play.

VI-B1. La política de privacidad permite mayor granularidad: La política de privacidad es un documento extenso que puede detallar con precisión qué datos específicos son recopilados, con información sobre el propósito de uso, el almacenamiento y el posible intercambio con terceros. Esto permite a los desarrolladores proporcionar una descripción detallada de los tipos de datos personales recopilados y su finalidad.

En la tabla I, se observa que algunos datos, como la *dirección MAC*, la *lista de amigos en Retos* y la *sincronización de contactos*, aparecen en la política de privacidad, pero no están reflejados en la seguridad de los datos en Google Play. Esta diferencia sugiere que la política de privacidad ofrece información más exhaustiva y detallada sobre la recopilación y tratamiento de datos.

VI-B2. Google Play prioriza la simplicidad en la información para los usuarios: La seguridad de los datos en Google Play, en cambio, parece estar más enfocada en ofrecer una representación simplificada de la información, agrupando los datos en categorías generales como:

- **Información de salud** en lugar de especificar si se trata de frecuencia cardíaca, presión arterial o nivel de oxígeno en sangre.
- **Información de actividad física** sin detallar si se trata de pasos, velocidad o distancia recorrida.

- **IDs de usuario** sin diferenciar entre identificadores de cuenta, número de teléfono o dirección MAC.

Esta falta de granularidad puede generar discrepancias con la política de privacidad. Es posible que los usuarios necesiten consultar la política de privacidad para obtener un desglose más detallado de los datos recopilados por la aplicación.

VI-C. Análisis cuantitativo de las discrepancias

A partir de los datos obtenidos en la Tabla VI, se ha realizado un análisis cuantitativo de las discrepancias entre las diferentes fuentes de información.

La Tabla IV resume los resultados. En ella se presentan el número total de permisos solicitados en el archivo `AndroidManifest.xml` relacionados con datos personales, los datos personales identificados y la cantidad de comparaciones realizadas entre las distintas fuentes. Además, se detallan las discrepancias detectadas en cada comparación, lo que permite evaluar el grado de coherencia entre las declaraciones de privacidad y los permisos técnicamente requeridos por la aplicación.

De acuerdo con los resultados, se identificaron un total de 50 permisos solicitados por la aplicación, mientras que los datos personales declarados alcanzaron un total de 54. En términos de comparaciones entre fuentes, se realizaron 78.

El análisis revela un alto número de discrepancias, con 20 inconsistencias entre la política de privacidad y Google Play, 21 entre Google Play y los permisos, y 37 entre la política de privacidad y los permisos.

Estos resultados evidencian que la falta de alineación entre las fuentes de información, lo que podría afectar la transparencia y la confianza de los usuarios en la aplicación.

VI-D. Porcentaje de discrepancias entre fuentes

Para dimensionar la magnitud de las discrepancias, se calcularon los porcentajes de inconsistencias entre las distintas fuentes de información, como se muestra en la Tabla V. Los

Tabla IV
RESUMEN DE PERMISOS, DATOS PERSONALES Y DISCREPANCIAS IDENTIFICADAS

Métrica	Total
Permisos	50
Datos personales	54
Comparaciones	78
PP ↔ GP	20
GP ↔ Permisos	21
PP ↔ Permisos	37

Tabla V
DISCREPANCIAS ENTRE FUENTES

Métrica	(%)
PP ↔ GP	25.64
GP ↔ Permisos	26.92
PP ↔ Permisos	47.44

Leyenda: ↔ indica discrepancia.

resultados indican que el 25.64 % de las comparaciones entre la política de privacidad y Google Play presentan discrepancias, lo que sugiere que la política de privacidad ofrece una descripción más detallada de los datos recopilados que la sección de seguridad de datos en Google Play. Esto se vé en la tabla I donde dentro de una misma categoría de Google play se engloban varios datos personales.

Por otro lado, el 26.92 % de las comparaciones entre Google Play y los permisos presentan inconsistencias, lo que evidencia que la sección de seguridad de datos en Google Play no refleja con precisión los permisos solicitados por la aplicación.

Finalmente, el 47.44 % de las comparaciones entre la política de privacidad y los permisos también muestran discrepancias, lo que refuerza la necesidad de una mayor alineación entre estas fuentes para mejorar la claridad y precisión de la información proporcionada a los usuarios.

Este análisis cuantitativo complementa los hallazgos previos y subraya la importancia de mejorar la coherencia entre las diferentes fuentes de información para garantizar una mayor transparencia en el manejo de datos personales por parte de Samsung Health.

VII. CONCLUSIÓN

El estudio realizado sobre la aplicación *Samsung Health* reveló discrepancias entre la política de privacidad, la sección de Seguridad de los Datos en Google Play y los permisos solicitados en el archivo `AndroidManifest.xml`. Se identificó que la mera presencia de permisos en Android no siempre implica acceso directo a datos personales, ya que algunos simplemente proporcionan acceso técnico a recursos del dispositivo. Sin embargo, esto genera confusión si la política de privacidad y la información en Google Play no especifican claramente el propósito exacto de estos permisos.

Además, los permisos específicos introducidos por la API Health Connect representan un avance significativo hacia una mayor transparencia, permitiendo un control más granular sobre los datos sensibles como el ritmo cardíaco o niveles de glucosa. No obstante, se evidenció que la política de privacidad proporciona un nivel de detalle superior al de Google Play, cuya simplificación facilita la comprensión general pero puede ocultar detalles relevantes, generando opacidad en la comunicación con los usuarios.

Finalmente, el análisis subraya la necesidad de una mejor alineación y precisión entre la documentación proporcionada por las aplicaciones y los permisos efectivamente solicitados.

VIII. TRABAJO FUTURO

Como trabajo futuro, se completará la automatización de todo el proceso, con especial énfasis en el análisis automático de las políticas de privacidad.

La automatización de este proceso no solo permitirá mejorar la eficiencia del análisis, sino que también facilitará su aplicación a un mayor número de aplicaciones, contribuyendo a una mayor transparencia en la gestión de datos personales dentro del ecosistema de aplicaciones móviles.

AGRADECIMIENTOS

Este trabajo se incluye en las actividades del Proyecto Estratégico de Ciberseguridad App-PI (*App Privacy Impact*): Un ecosistema para la evaluación del impacto de apps para dispositivos móviles sobre la privacidad y seguridad de sus usuarios, el cual se realiza al amparo de un convenio de colaboración entre la Universidad de Valladolid y la S.M.E. Instituto Nacional de Ciberseguridad de España M.P., S.A. para la promoción de proyectos estratégicos de Ciberseguridad en España, en el marco de los Fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation).

Tabla VI
FRAGMENTO DE LA TABLA DE ALINEAMIENTOS.
ANÁLISIS DE DISCREPANCIAS Y SIMILITUDES ENTRE PERMISOS, POLÍTICA DE PRIVACIDAD Y SEGURIDAD DE DATOS EN GOOGLE PLAY PARA SAMSUNG HEALTH VERSIÓN 6.29.2.001

Nº	Permiso	Dato Personal	PP	GP	PP ↔ GP	GP ↔ P	PP ↔ P
1	WRITE_EXTERNAL_STORAGE	Archivos personales	No	Sí	No	Sí	No
2	READ_EXTERNAL_STORAGE	Archivos personales	No	Sí	No	Sí	No
3	READ_MEDIA_IMAGES	Imágenes personales	No	Sí	No	Sí	No
5	ACCESS_FINE_LOCATION	Ubicación precisa	Sí	Sí	Sí	Sí	Sí
6	ACCESS_COARSE_LOCATION	Ubicación aproximada	No	No	Sí	No	No
8	BODY_SENSORS	Datos biométricos	Sí	Sí	Sí	Sí	Sí
9	CAMERA	Imágenes personales	No	Sí	No	Sí	No
10	READ_CONTACTS	Contactos	Sí	Sí	Sí	Sí	Sí
11	READ_PHONE_NUMBERS	Número de teléfono	Sí	Sí	Sí	Sí	Sí
12	READ_PHONE_STATE	Estado del dispositivo	No	No	Sí	No	No
13	GET_ACCOUNTS	Cuentas del usuario	Sí	Sí	Sí	Sí	Sí
15	RECORD_AUDIO	Voz del usuario	No	Sí	No	Sí	No
16	HIGH_SAMPLING_RATE_SENSORS	Datos biométricos	Sí	Sí	Sí	Sí	Sí
17	READ_HEART_RATE	Frecuencia cardíaca	Sí	Sí	Sí	Sí	Sí
18	WRITE_HEART_RATE	Frecuencia cardíaca	Sí	Sí	Sí	Sí	Sí
35	READ_TOTAL_CALORIES_BURNED	Calorías quemadas	No	No	Sí	No	No
36	WRITE_TOTAL_CALORIES_BURNED	Calorías quemadas	No	No	Sí	No	No
39	READ_VO2_MAX	VO2 máximo	No	No	Sí	No	No
40	WRITE_VO2_MAX	VO2 máximo	No	No	Sí	No	No
43	READ_POWER	Potencia de ejercicio	No	No	Sí	No	No
44	WRITE_POWER	Potencia de ejercicio	No	No	Sí	No	No
47	READ_NUTRITION	Datos de nutrición	No	No	Sí	No	No
48	WRITE_NUTRITION	Datos de nutrición	No	No	Sí	No	No
51	GET_ACCOUNTS	ID de Samsung account	Sí	Sí	Sí	Sí	Sí
52	–	Fecha de nacimiento	Sí	No	No	Sí	No
53	READ_MEDIA_IMAGES, CAMERA	Foto de perfil	Sí	Sí	Sí	Sí	Sí
54	GET_ACCOUNTS	Correo electrónico	Sí	No	No	No	Sí
55	–	Código postal	Sí	No	No	Sí	No
56	–	Publicaciones	Sí	No	No	Sí	No
57	–	País	Sí	No	No	Sí	No
58	BODY_SENSORS	Músculo esquelético	Sí	Sí	Sí	Sí	Sí
59	–	Información del ciclo menstrual	Sí	Sí	Sí	No	No
60	–	Consumo de agua	Sí	Sí	Sí	No	No
61	–	Mensajes enviados en Retos	Sí	No	No	Sí	No
62	–	Dirección MAC	Sí	No	No	Sí	No
63	–	Identificadores de publicidad móvil	Sí	Sí	Sí	No	No
64	–	Diagnósticos	Sí	Sí	Sí	No	No
65	–	Lista de amigos en Retos	Sí	No	No	Sí	No
66	GET_ACCOUNTS	Nombre de usuario	Sí	Sí	Sí	Sí	Sí
67	–	Sexo	Sí	No	No	Sí	No
68	–	Publicaciones	Sí	No	No	Sí	No
69	BODY_SENSORS	Masa muscular	Sí	Sí	Sí	Sí	Sí
70	READ_HEART_RATE, HIGH_SAMPLING_RATE_SENSORS, BODY_SENSORS	Variabilidad de la frecuencia cardíaca	Sí	Sí	Sí	Sí	Sí
71	–	Estado de ánimo	Sí	Sí	Sí	No	No
72	–	Datos de dispositivos vinculados	Sí	Sí	Sí	No	No
73	–	Interacciones con la aplicación	Sí	Sí	Sí	No	No
74	–	Señales Bluetooth y Wi-Fi	Sí	No	No	Sí	No
75	READ_PHONE_STATE	Número de serie del dispositivo	Sí	No	No	No	Sí
76	–	Registros de fallos	Sí	Sí	Sí	No	No
77	–	Historial de sincronización de contactos	Sí	Sí	Sí	No	No
78	WRITE_EXTERNAL_STORAGE, READ_EXTERNAL_STORAGE	Archivos y documentos	No	Sí	No	Sí	No

Leyenda:

- **Permiso (P)**: Permisos solicitados por la aplicación en el archivo `AndroidManifest.xml`.
- **Dato Personal**: Tipo de información personal asociada al permiso solicitado o mencionada en la política o Google Play.
- **PP (Política de Privacidad)**: **Sí** indica que el dato personal está declarado en la política de privacidad; **No** indica ausencia en la política.
- **GP**: El dato personal aparece en la sección Seguridad de Datos de Google Play (**Sí**) o no aparece (**No**).
- **PP ↔ GP**: **Sí** indica coherencia entre política y Google Play; **No** indica discrepancia.
- **GP ↔ Permisos**: **Sí** indica coherencia entre permisos solicitados y lo declarado en Google Play; **No** indica discrepancia.
- **PP ↔ Permisos**: **Sí** indica coherencia entre permisos solicitados y la política de privacidad; **No** indica discrepancia.
- **–**: Indica que no aplica o no existe información relacionada.

REFERENCIAS

[1] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj

[2] Agencia Española de Protección de Datos (AEPD), "El deber de informar y otras medidas de responsabilidad proactiva en apps para dispositivos móviles," Agencia Española de Protección de Datos, Tech. Rep., septiembre 2019. [Online]. Available: https://www.aepd.es/sites/default/files/2019-11/nota-tecnica-apps-moviles.pdf

[3] Commission Nationale Informatique & Libertés (CNIL), "Recommendation relative aux applications mobiles," Commission Nationale Informatique & Libertés, Tech. Rep., enero 2024. [Online]. Available: https://www.cnil.fr/sites/cnil/files/2024-09/recommandation-applications-mobiles.pdf

[4] Google Play Console Help, "Data safety section in google play - play console help," 2025, last accessed: March 2, 2025. [Online]. Available: https://support.google.com/googleplay/android-developer/answer/10787469

[5] M. Liu and Q. Wang, "Research on android user privacy permission analysis and protection mechanism under big data environment," *MATEC Web of Conferences*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269854220

[6] M. M. Martínez-González, A. Pérez-Fuente, A. Aparicio, and P. A. Criado-Lozano, "Using the metadata-based app-pi ecosystem to assess the privacy impact of health apps," in *UCAmI 2024 - International Conference on Ubiquitous Computing and Ambient Intelligence*. Springer Nature Switzerland AG, 2024, pp. 522–533. [Online]. Available: https://doi.org/10.1007/978-3-031-77571-0_50

[7] A. Papageorgiou, M. Strigkos, E. A. Politou, E. Alepis, A. Solanas, and C. Patsakis, "Security and privacy analysis of mobile health applications: The alarming state of practice," *IEEE Access*, vol. 6, pp. 9390–9403, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3878450

[8] Google Play Console, "Data safety section for google play apps," 2024, accessed: 2024-02-23. [Online]. Available: https://support.google.com/googleplay/android-developer/table/12921780?hl=en

[9] Google Play Console Help, "Requisitos de datos de salud en google play," 2025, Último acceso: 2 de marzo de 2025. [Online]. Available: https://support.google.com/googleplay/android-developer/answer/12261419?sjid=6948904469239438040-EU

[10] W. Enck, P. Gilbert, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones," in *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation (OSDI)*, 2010. [Online]. Available: https://www.usenix.org/legacy/event/osdi10/tech/full_papers/Enck.pdf

[11] R. Slavin, X. Wang, M. B. Hosseini, J. Hester, R. Krishnan, J. Bhatia, T. D. Breaux, and J. Niu, *ICSE 2016 : 2016 IEEE/ACM 38th IEEE International Conference on Software Engineering : proceedings : 14-22 May 2016, Austin, Texas, USA*. IEEE Computer Society : The Association for Computing Machinery, 2016. [Online]. Available: https://www.cs.cmu.edu/~breaux/publications/slavin-icse16.pdf

[12] S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Schaub, S. Wilson, N. Sadeh, S. M. Bellovin, and J. Reidenberg, "Automated analysis of privacy requirements for mobile apps," in *24th Annual Network and Distributed System Security Symposium, NDSS 2017*. The Internet Society, 2017. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2017/09/ndss2017_05A-5_Zimmeck_paper.pdf

[13] L. Yu, X. Luo, X. Liu, and Z. Tao, "Can we trust the privacy policies of android apps?" *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 538–549, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:1922862

[14] G. Tangari, M. Ikram, K. Ijaz, M. A. Kaafar, and S. Berkovsky, "Mobile health and privacy: cross sectional study," *BMJ*, jun 2021.

[15] J. Benjumea, J. Ropero, O. Rivera-Romero, E. Dorronzoro-Zubiete, and A. Carrasco, "Privacy assessment in mobile health apps: Scoping review," *JMIR Mhealth Unhealth*, vol. 8, no. 7, 2020.

[16] G. Tangari, M. Ikram, I. Wayan, B. Sentana, K. Ijaz, M. A. Kâafar, and S. Berkovsky, "Analyzing security issues of android mobile health and medical applications," *Journal of the American Medical Informatics Association : JAMIA*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236776061

[17] Android Developers, "Tipos de datos en health connect - android developers," 2025, last accessed: March 2, 2025. [Online]. Available: https://developer.android.com/health-and-fitness/guides/health-connect/plan/data-types?hl=es-419

[18] "Zenodo," https://zenodo.org/records/14996871.

[19] Samsung, "Privacy — samsung health," https://www.samsunghealth.com/privacy/, 2025, accessed: March 17, 2025.

¿Es *Health Connect* la mejor garantía de Seguridad y Privacidad para los datos de salud?

Alejandro Pérez-Fuente
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid, España
alejandro.perez.fuente@uva.es

M. Mercedes Martínez-González
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid, España
mercedes@infor.uva.es

Amador Aparicio
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid, España
amador@infor.uva.es

Pablo A. Criado-Lozano
Grupo de Ingeniería de la Privacidad (IngPriv)
Universidad de Valladolid, España
Universidad Europea Miguel de Cervantes, España
pacriado@uemc.es

***Resumen*—Ante la creciente preocupación por la seguridad y privacidad de los datos de salud, con la llegada de Android 14, se introduce *Health Connect*, una plataforma que busca mejorar su gestión y permite a los usuarios un mayor control.**

Pero, ¿es *Health Connect* una mejora en seguridad y privacidad frente a las soluciones utilizadas previamente? En este trabajo se responde a esta pregunta mediante un análisis novedoso que combina tanto los criterios de seguridad como los de privacidad emanados de la normativa. El análisis incluye su comparación frente a soluciones pasadas, como Google Fit y la API de Sensores de Android.

Los resultados muestran que *Health Connect* es la solución que aporta mejor protección de la información mediante almacenamiento local, cifrado y permisos más específicos. Sin embargo, aún debe abordar desafíos como la falta de transparencia en la documentación, que pueden afectar su adopción y confiabilidad.

***Index Terms*—Seguridad, Privacidad, Aplicaciones Móviles, Metadatos, Salud, Android**

Tipo de contribución: *Investigación original (límite 8 páginas)*

I. Introducción

La seguridad y la privacidad en las aplicaciones de salud son temas de creciente importancia [1], [2]. Con el aumento del uso de dispositivos móviles y plataformas tecnológicas para el monitoreo y manejo de la salud, los usuarios comparten constantemente datos sensibles, como información médica personal y biométrica [3]. Esta recopilación y el constante intercambio de información de salud entre distintas aplicaciones presentan riesgos significativos, ya que, si no se protegen adecuadamente, estos datos podrían quedar expuestos a usos indebidos. Por tanto, desarrollar mecanismos efectivos de protección y control de acceso es crucial para salvaguardar la integridad de los sistemas donde se intercambian datos de salud y evitar posibles brechas de seguridad y privacidad.

En este contexto, Android, el sistema operativo más utilizado en dispositivos móviles [4], juega un papel crucial en el ecosistema de las aplicaciones de salud. Tanto los *wearables* como las aplicaciones de salud se conectan y funcionan con Android. Por tanto, en última instancia, la seguridad de los datos sensibles depende de lo seguro que sea este sistema operativo. La creciente preocupación por la seguridad y la privacidad en los dispositivos móviles radica, en parte, en la dificultad de proteger los datos personales de manera específica para cada aplicación, lo que expone a los usuarios. Hasta ahora, Android ha presentado una notable falta de granularidad en el control de acceso a los recursos y datos sensibles del dispositivo [5].

Poco a poco, Android ha ido implementando cambios que han transformado la manera en que las aplicaciones acceden, almacenan y comparten información personal[1]. Uno de los avances más significativos es la introducción de *Health Connect*, anunciado en la Google I/O de 2022[2] como una plataforma destinada a redefinir la gestión de los datos de salud en Android. Esta plataforma comenzó a operar con la llegada de Android 14 en octubre de 2023.

Se ha planteado la cuestión que motiva este estudio: ¿Supone *Health Connect* un avance significativo en términos de seguridad y privacidad frente a las soluciones que la preceden? Más aún: ¿en ambos, quizá solo uno de ellos, o incluso en ninguno de ellos? Para responder a estas preguntas, se ha llevado a cabo un análisis innovador y detallado sobre la repercusión de *Health Connect* en la seguridad y privacidad de la información de salud en Android 14, destacando su impacto frente a soluciones anteriores. Se ha diseñado un método de análisis que se estructura en dos fases complementarias que permiten una comprensión integral de la repercusión de esta tecnología.

En la primera fase, se examinan en profundidad las funcionalidades de *Health Connect* en el contexto del sistema operativo Android 14. Se ha puesto especial atención en los mecanismos de seguridad de los datos y de gestión del acceso a la información de salud. Este análisis es imprescindible para comprender el funcionamiento de la plataforma y detectar los aspectos que pueden influir en la seguridad y privacidad de los usuarios. En la segunda fase, se hace un análisis sistemático guiado por los requisitos de seguridad y privacidad, evaluando hasta qué punto *Health Connect* satisface estos requisitos en comparación con soluciones anteriores. Este enfoque permite valorar la calidad de *Health Connect* y responder a las preguntas que han motivado este estudio.

[1]https://support.google.com/googleplay/android-developer/answer/9888170

[2]https://hackernoon.com/google-io-2022-key-announcements-for-developers

El resto del artículo se organiza de la siguiente manera. En la sección II, se revisa el estado del arte sobre la gestión de recursos y datos sensibles en Android. En la sección III, se analizan los estudios previos relacionados con la seguridad y privacidad de los datos de salud, así como el almacenamiento seguro, el cifrado y las normativas aplicables. En la sección IV, se examinan los cambios introducidos por *Health Connect* en los mecanismos de protección de datos y de gestión del acceso a la información en Android. En la sección V, se estudian las repercusiones de estos cambios sobre la seguridad y la privacidad en comparación con otras soluciones. En la sección VI, se analizan los desafíos asociados a la adopción de *Health Connect* en Android. En la sección VII, se presentan las conclusiones. Finalmente, en la sección VIII, se proponen posibles líneas de trabajo futuro.

II. Estado del Arte

El componente del sistema operativo Android responsable de gestionar el acceso a los recursos y datos sensibles del dispositivo es el modelo de seguridad de Android. De manera más específica, es el modelo de permisos, que constituye una parte fundamental de este sistema de seguridad.

Con el fin de mejorar la seguridad y privacidad de los usuarios, el modelo de permisos de Android ha experimentado una evolución significativa desde su introducción en 2008. En sus primeros años, las aplicaciones solicitaban permisos únicamente durante la instalación, otorgando acceso completo a diversas funcionalidades del dispositivo. Con la llegada de Android 6, se implementó un modelo de permisos en tiempo de ejecución, lo que permitió a los usuarios otorgar o revocar permisos según sus necesidades [6]. Más tarde, con Android 11, se introdujeron permisos más granulares para el acceso a la ubicación, el micrófono y la cámara, ofreciendo a los usuarios un control más detallado sobre qué datos y funcionalidades podían ser accedidos por las aplicaciones. Además, Android 13 continuó esta evolución al incorporar permisos específicos para diferentes tipos de archivos multimedia, como imágenes, vídeos y audios, reemplazando el permiso general de almacenamiento externo [7].

Estos cambios reflejan el esfuerzo continuo de Google por ofrecer a los usuarios un control más preciso sobre los datos y funcionalidades a los que las aplicaciones pueden acceder. En principio, estas modificaciones buscan mejorar la experiencia de los usuarios, reforzando su seguridad y privacidad. El cambio más reciente en esta línea es la introducción de *Health Connect*, que se aborda en las secciones IV y V.

III. Trabajo Relacionado

El cumplimiento de normativas como el Reglamento General de Protección de Datos (RGPD) en Europa [8] o la Ley de Portabilidad y Responsabilidad del Seguro Médico (HIPAA) en EE. UU. [9] es obligatorio y fundamental para garantizar la seguridad y privacidad de los datos en las aplicaciones de salud. Sin embargo, diversos estudios han demostrado que muchas de estas aplicaciones no cumplen completamente con dichas regulaciones, ya sea por falta de transparencia en el manejo de los datos [10], [11], [12], ausencia de mecanismos adecuados de gestión del acceso a los datos o deficiencias en las medidas de seguridad implementadas [13], [10], [11].

En numerosos casos, el cifrado de los datos es insuficiente, lo que puede dar lugar a filtraciones de datos personales, incluidos datos sensibles como credenciales, localización y otros, que se transmiten a través del protocolo HTTP sin cifrado en tránsito [14], [15]. De hecho, diversas investigaciones han demostrado que las aplicaciones que emplean protocolos inseguros o que almacenan información sensible sin cifrar son más vulnerables a ataques por parte de actores malintencionados, lo que facilita las filtraciones de datos [16], [17]. Sin embargo, estos estudios no abordan el caso de *Health Connect*, dado que su implementación es reciente, lo que crea un vacío en la literatura respecto a su impacto en la seguridad y privacidad en la gestión de datos en aplicaciones móviles.

Como se mencionó en la sección I, el modelo de seguridad de Android presenta un problema de granularidad en la gestión del acceso a recursos y datos sensibles. Esto afecta especialmente a las aplicaciones de salud, ya que los permisos que solicitan proporcionan acceso a una amplia gama de datos con un solo permiso, como `BODY_SENSORS`, que da acceso a los datos de sensores de salud, o `READ_EXTERNAL_STORAGE`, que permite acceder a cuantos datos se almacenen en el almacenamiento externo del dispositivo [18]. Este enfoque puede comprometer el cumplimiento de normativas como el Reglamento General de Protección de Datos, que exige un acceso más restringido y controlado a los datos personales, así como la obtención de un consentimiento explícito para cada tipo de dato [18], [16]. Además, la falta de granularidad en los permisos incrementa el riesgo de violaciones del RGPD [16], lo que resulta en riesgos para la privacidad de los usuarios.

IV. Subiendo la granularidad en Android

Como se ha mencionado en la sección II, Android ha refinado progresivamente su modelo de permisos para ofrecer a los usuarios un control más preciso y detallado sobre los recursos de sus dispositivos. A continuación, se analizan los cambios introducidos en Android 14, con especial énfasis en los mecanismos de seguridad de los datos y de gestión del acceso a la información de salud. Este análisis es fundamental para evaluar su impacto en la seguridad y privacidad de la información de salud del usuario.

IV-A. Evolución del recurso al dato en Android 14

En versiones anteriores a Android 14, las aplicaciones que requieren acceso a datos de salud utilizan la API (*Application Programming Interface*) de sensores de Android[3]. Los sensores *hardware*, como el podómetro o el sensor de frecuencia cardíaca, pueden consultarse a través de la clase `SensorManager`, siempre que la aplicación cuente con el permiso `BODY_SENSORS` [19]. Una vez concedido este permiso, las aplicaciones pueden acceder de manera continua a los datos sin necesidad de solicitar confirmaciones adicionales al usuario.

Este modelo de acceso presenta varias limitaciones. Por ejemplo, si una aplicación solo necesita un dato específico, como la presión arterial, debe solicitar el permiso `BODY_SENSORS`, el cual otorga acceso a todos los sensores de salud del dispositivo, incluso a aquellos que la aplicación

[3]https://developer.android.com/reference/android/hardware/Sensor

no requiere[4]. Esta falta de granularidad en la gestión del acceso a datos ha generado preocupaciones sobre la seguridad, la privacidad y el control del usuario [14], [20].

Una alternativa para gestionar los datos de salud es la sincronización con Google Fit, un servicio basado en computación en la nube que permite a las aplicaciones acceder a esta información a través de su plataforma. Aunque se trata de una solución ampliamente utilizada, dejará de estar disponible a partir del 30 de junio de 2025[5]. Este enfoque conlleva riesgos inherentes a la computación en la nube, especialmente en términos de seguridad y control de acceso [21]. Al depender de un servicio externo, los datos de salud deben transmitirse y almacenarse fuera del dispositivo, lo que amplía la superficie de ataque en comparación con el acceso directo a los sensores. Esta externalización expone la información a posibles vulnerabilidades en la infraestructura del proveedor, aumentando el riesgo de accesos no autorizados o filtraciones de datos [22]. Por tanto, los usuarios deben confiar en cómo Google gestiona todos estos riesgos. Además, el almacenamiento en la nube limita el control directo por parte de los usuarios sobre el acceso a sus propios datos, a diferencia de una solución local, donde la información permanece en el dispositivo y bajo el dominio exclusivo del usuario.

Finalmente, desde Android 14, *Health Connect* supone un cambio de modelo en el control de acceso en el sistema operativo. Donde antes el único elemento sobre el que se otorgaban permisos era el recurso, ahora se introduce la posibilidad de solicitar permiso para acceder a un dato concreto. Este cambio, cuya trascendencia podría pasar desapercibida para los desarrolladores poco familiarizados con el reto que supone cumplir adecuadamente con los estrictos requisitos de privacidad derivados de la normativa en vigor, es en realidad un cambio sustancial: el nivel de granularidad con el que se puede trabajar aumenta. Ya no se tiene acceso al conjunto de datos de salud en bloque. Ahora se accede a cada dato de salud individualmente.

IV-B. Health Connect

Health Connect es una solución creada por Android para centralizar y gestionar la información de salud localmente en el dispositivo del usuario. Su objetivo es unificar el acceso a esta información, facilitando la interoperabilidad entre diversas aplicaciones mediante un modelo seguro y eficiente de intercambio de datos. A partir del 28 de mayo de 2025[6], la implementación de *Health Connect* será obligatoria para los desarrolladores de aplicaciones de salud en Android. *Health Connect* se compone de varios elementos clave, entre los cuales destacan: una API de sincronización y una aplicación de gestión.

IV-B1. API de sincronización: Este componente permite a los desarrolladores acceder al almacén de datos de salud de *Health Connect*, actuando como un repositorio intermedio que facilita la integración entre distintas aplicaciones. Su objetivo es garantizar que la información se mantenga actualizada y accesible únicamente para las aplicaciones autorizadas, mejorando así la interoperabilidad y la seguridad de los datos. Desde octubre de 2023, con el lanzamiento de Android 14, esta API está integrada en el sistema operativo como parte de la biblioteca *Jetpack Health*[7], proporcionando una solución nativa y más eficiente para la gestión de datos de salud.

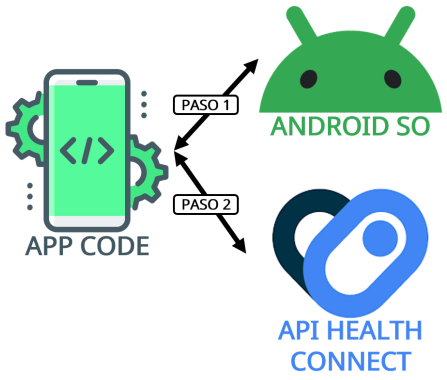


Figura 1. Interacción de una aplicación con Android y *Heatlh Connect*

En la figura 1 se ilustra cómo una aplicación interactúa con la API de *Health Connect* para su funcionamiento. A diferencia de otras soluciones, como la API de Sensores de Android o Google Fit, donde la aplicación interactúa directamente con un solo componente, la implementación de la API de *Health Connect* requiere un acceso en dos etapas. Como preparativos a la interacción, la aplicación debe declarar y solicitar los permisos necesarios dentro del archivo `AndroidManifest.xml` para acceder a los datos específicos, asegurándose también de que el sistema operativo en el que se ejecuta sea compatible con *Health Connect*, lo que implica añadir su SDK (*Software Development Kit*) como dependencia en el archivo `build.gradle`. En el siguiente recuadro se muestran dos fragmentos de código extraídos de dos de estos ficheros, `AndroidManifest.xml` y `build.gradle`.

App AndroidManifest.xml

```
<uses-permission android:name="android.
  permission.health.READ_HEART_RATE"/>
```

App build.gradle

```
dependencies {
  implementation "androidx.health.connect
          :connect-client:1.1.0-alpha11"
}
```

En el primer paso, la aplicación solicita acceso al componente *Health Connect* a través del punto de entrada `HealthConnectClient` del sistema operativo. Una vez que se obtiene el componente y se conceden los permisos necesarios, en el segundo paso, la aplicación interactúa con la API de *Health Connect* y realiza las operaciones requeridas.

Una de las características clave de seguridad de *Health Connect* es que dentro de su almacén, según Google, los datos se encuentran cifrados[8]. Sin embargo, la documentación

[4]https://developer.android.com/reference/android/Manifest.permission#BODY_SENSORS

[5]https://developers.google.com/fit/android

[6]https://support.google.com/googleplay/android-developer/table/12921780?hl=en

[7]https://developer.android.com/jetpack/androidx/releases/health

[8]https://developer.android.com/codelabs/health-connect#0

HealthConnectClient.kt

```
60
61      /**
62       * Inserts one or more [Record] and returns newly assigned
63       * [androidx.health.connect.client.records.metadata.Metadata.id] generated. Insertion of
64       * multiple [records] is executed in a transaction - if one fails, none is inserted.
65       *
66       * @param records List of records to insert
67       * @return List of unique identifiers in the order of inserted records.
68       * @throws android.os.RemoteException For any IPC transportation failures.
69       * @throws SecurityException For requests with unpermitted access.
70       * @throws java.io.IOException For any disk I/O issues.
71       *
72       * For example, to insert basic data like step counts:
73       *
74       * @sample androidx.health.connect.client.samples.InsertSteps
75       *
76       * To insert more complex data like nutrition for a user who's eaten a banana:
77       *
78       * @sample androidx.health.connect.client.samples.InsertNutrition
79       *
80       * To insert some heart rate data:
81       *
82       * @sample androidx.health.connect.client.samples.InsertHeartRateSeries
83       *
84       * [androidx.health.connect.client.records.metadata.Metadata.clientRecordId] can be used to
85       * deduplicate data with a client provided unique identifier. When a subsequent [insertRecords]
86       * is called with the same
87       * [androidx.health.connect.client.records.metadata.Metadata.clientRecordId], whichever [Record]
88       * with the higher
89       * [androidx.health.connect.client.records.metadata.Metadata.clientRecordVersion] takes
90       * precedence.
91       */
92      suspend fun insertRecords(records: List<Record>): InsertRecordsResponse
```

Figura 2. Almacenamiento en *Health Connect*

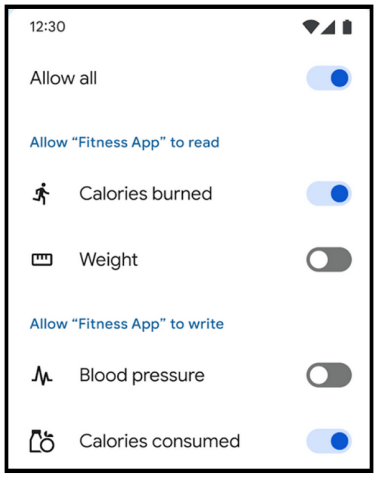


Figura 3. Gestión de datos dentro de la aplicación *Health Connect*.

oficial no especifica ni el algoritmo de cifrado utilizado ni el mecanismo de gestión de las claves.

Para descubrir su funcionamiento, se ha analizado el código de *Health Connect*. En la Figura 2 se muestra un fragmento del código de *Health Connect* en el que se realiza el almacenamiento de un registro de salud. En este, se observa cómo el registro sin cifrar se envía para su almacenamiento a través de un mensaje que ingresa en Android IPC (*Inter-Process Communication*). Por lo tanto, el proceso responsable de almacenar y cifrar los datos es otro componente del sistema operativo Android.

Este cifrado, junto con el acceso en dos pasos, en nuestra opinión, puede ser una fuente de mayor seguridad y privacidad, como argumentaremos en la sección V.

IV-B2. Aplicación de gestión: Este componente ofrece a los usuarios un control minucioso sobre el acceso a sus datos de salud almacenados en *Health Connect*. Permite gestionar de forma precisa qué aplicaciones pueden acceder a cada tipo de dato y definir si dicho acceso será de solo lectura o incluirá también la escritura, mediante la concesión o revocación de permisos. Se incorpora de manera nativa en los dispositivos con Android 14 y está disponible para su descarga en Play Store.

La figura 3 muestra cómo gestionar los permisos de una aplicación hipotética llamada “Fitness App” a través de *Health Connect*. Como se puede observar, la aplicación de gestión ofrece un acceso en bloque a todos los datos de salud mediante una opción de “Permitir todo”. Sin embargo, si los usuarios desean un control más detallado, pueden gestionar los permisos de forma granular mediante dos listas separadas: una para los permisos de lectura de datos y otra para los de escritura. En la figura, la primera lista corresponde a los datos a los que la aplicación tiene permiso para acceder en modo lectura, en este caso, las calorías quemadas. La segunda lista muestra los datos que la aplicación puede escribir, que en este ejemplo son las calorías ingeridas.

El nivel de granularidad en el control del acceso a los datos de salud ha aumentado notablemente respecto a las soluciones anteriores. En la sección V se discute su efecto sobre los dos aspectos que nos interesan: seguridad y privacidad.

V. Repercusión de Health Connect sobre la seguridad y privacidad

Al inicio de la sección IV se detectaron lo que podrían ser algunas de las posibles alternativas a *Health Connect*, entre las que destacan la API de Sensores de Android y Google Fit. Posteriormente, en esa misma sección, al analizar qué es y cómo funciona *Health Connect*, se señalaron ciertas características clave que podrían repercutir en la seguridad y la privacidad, como el acceso en dos pasos al almacén, el cifrado de los datos y la mayor granularidad en la gestión del acceso a la información.

Partiendo de las distintas alternativas identificadas y de las características clave, se plantea un análisis comparativo de la seguridad y la privacidad de *Health Connect* frente a sus alternativas, basándose en la documentación oficial de cada uno de estos sistemas. En primer lugar, se definen los criterios de comparación a utilizar. A continuación, se elabora una tabla que aplica estos criterios a cada una de las alternativas. Finalmente, se analizan los resultados obtenidos y se examinan otras posibles repercusiones en materia de seguridad y privacidad. De esta forma, se obtiene una visión clara de la calidad de *Health Connect* y de su contribución a la mejora de la seguridad y privacidad de los datos de salud en el ecosistema Android.

V-A. Seguridad en Health Connect

La seguridad informática se sustenta en varios pilares esenciales que protegen la información en entornos digitales[9]. De entre estos, analizaremos aquellos que están relacionados con los datos de salud.

Confidencialidad Garantiza que los datos sean accesibles únicamente para quienes cuentan con la autorización correspondiente, evitando filtraciones o accesos indebidos.

Integridad Asegura que la información se mantenga exacta y libre de modificaciones no autorizadas, preservando su veracidad y consistencia. Es especialmente crucial garantizarla durante la transmisión de los datos.

[9]https://unirfp.unir.net/revista/ingenieria-y-tecnologia/principios-seguridad-informatica/

Disponibilidad Se encarga de que tanto los sistemas como los datos estén a disposición de los usuarios autorizados cuando se requieran, evitando interrupciones o bloqueos en el servicio.

Autenticidad Verifica la identidad de los usuarios y la procedencia de la información, estableciendo una base sólida de confianza en las comunicaciones y transacciones digitales.

Estos pilares son esenciales para evaluar la seguridad de cada propuesta. No obstante, garantizar su cumplimiento absoluto en cada solución es un desafío complejo que excede el alcance de este análisis. En su lugar, se evalúa en qué medida cada solución logra aproximarse a la garantía de cada pilar bajo ciertas circunstancias específicas. Se parte de la base que, como usuarios o desarrolladores, solo tenemos control sobre lo que ocurre localmente. Además, se asume que ni el dispositivo ni el sistema operativo están comprometidos.

Teniendo en cuenta estas consideraciones, la tabla I presenta una comparativa de seguridad entre *Health Connect*, Google Fit y la API de Sensores de Android. En cada fila se presenta la evaluación de cada criterio (Pilar de Seguridad) para las diferentes soluciones analizadas: *Health Connect*, Google Fit y la API de Sensores de Android. Para cada una de ellas se indica si es capaz de aportar a los desarrolladores mecanismos suficientes para garantizarla (Sí/No) y las razones que justifican este valor. Un "Sí" significa que los desarrolladores tienen suficientes instrumentos para ello, mientras que un "No" corresponde a la situación en que los desarrolladores se ven obligados a buscar mecanismos externos/ajenos a la solución estudiada para alcanzar ese criterio de seguridad. Por ejemplo, Google Fit por si mismo no es capaz de garantizar la confidencialidad. Una de las razones es la operación remota que supone el uso de la nube.

Evaluación del cumplimiento de cada uno de los pilares de seguridad seleccionados por *Health Connect*:

- *Confidencialidad: Health Connect* garantiza un alto nivel de confidencialidad gracias a su implementación local, que dificulta la interceptación de datos por parte de actores externos. Además, su almacén de datos está cifrado, lo que protege la información incluso en caso de accesos no autorizados a nivel local. Su sistema de gestión de acceso, altamente granular, permite un control preciso sobre qué aplicaciones pueden acceder a los datos, reforzando aún más la confidencialidad. Este aspecto es abordado dentro del modelo de seguridad de Android [6].
- *Integridad: Health Connect* garantiza la integridad de los datos al realizar los accesos de manera local, evitando modificaciones durante la transmisión. Además, los mecanismos de protección implementados en el sistema restringen los accesos no autorizados, impidiendo así alteraciones indebidas en la información almacenada [6].
- *Disponibilidad: Health Connect* garantiza una alta disponibilidad de los datos al operar de manera local en el dispositivo. Excepto en casos de fallos de *hardware* que afecten la memoria, la información almacenada en *Health Connect* permanece siempre accesible, sin depender de la conectividad a internet ni de servicios externos.
- *Autenticidad: Health Connect* garantiza la autenticidad de los datos a través de un sistema granular de gestión de accesos, que permite identificar qué aplicaciones están autorizadas para introducir información. De este modo, se asegura que los datos provienen de fuentes legítimas y confiables.

Evaluación del cumplimiento de cada uno de los pilares de seguridad seleccionados por Google Fit:

- *Confidencialidad:* En Google Fit, la confidencialidad depende en gran medida de la gestión de datos por parte de Google, ya que se trata de un servicio en la nube. A diferencia de *Health Connect*, el acceso a los datos se basa en permisos generales, lo que permite que cualquier aplicación con autorización pueda visualizar toda la información almacenada en la plataforma. Esta falta de granularidad en la gestión de permisos dificulta garantizar la confidencialidad a nivel individual.
- *Integridad:* La integridad de los datos en Google Fit no puede garantizarse plenamente, ya que, al ser una solución basada en la nube, está expuesta a vulnerabilidades que podrían comprometer la veracidad y consistencia de la información [22]. Factores como fallos de seguridad en los servidores o accesos no autorizados pueden afectar la fiabilidad de los datos almacenados.
- *Disponibilidad:* En Google Fit, la disponibilidad de los datos no está completamente garantizada, ya que al depender de la nube, puede verse afectada por interrupciones en la conectividad o fallos en los servidores de Google [23]. Problemas de red o mantenimiento de la plataforma pueden limitar el acceso a la información cuando se requiera.
- *Autenticidad:* Google Fit cumple con el criterio de autenticidad al incorporar un mecanismo que indica la fuente de los datos[10]. Este sistema permite verificar el origen de la información almacenada en la plataforma, asegurando su trazabilidad y fiabilidad.

Evaluación del cumplimiento de cada uno de los pilares de seguridad seleccionados por la API de Sensores de Android:

- *Confidencialidad:* La API de Sensores de Android no proporciona mecanismos para garantizar la confidencialidad de los datos recopilados. Al no contar con un sistema de gestión de acceso granular, cualquier aplicación con los permisos adecuados puede acceder a la información de los sensores sin restricciones específicas. Esta falta de control detallado impide garantizar la confidencialidad de los datos registrados.
- *Integridad:* La API de Sensores de Android asegura la integridad de la información, dado que los accesos se realizan localmente, eliminando el riesgo de modificaciones durante la transmisión. Además, el acceso a los sensores está protegido, lo que evita que aplicaciones no autorizadas puedan alterar los datos capturados.
- *Disponibilidad:* La API de Sensores de Android ofrece una disponibilidad total siempre que el dispositivo funcione correctamente. Dado que el acceso a los datos se realiza localmente, solo fallos de *hardware*, como problemas en los sensores o en la memoria del dispositivo, podrían impedir la recolección y consulta de la información en tiempo real.

[10]https://developers.google.com/fit/android/data-attribution

	Soluciones					
Pilar de Seguridad	***Health Connect***		**Google Fit**		**API de Sensores de Android**	
Confidencialidad	**Sí.**	Operación local. Permisos granulares. Almacenamiento cifrado. Contemplado en el modelo de seguridad de Android.	**No.**	Basada en la nube. Permisos generales.	**No.**	Permiso general.
Integridad	**Sí.**	Operación local. Contemplado en el modelo de seguridad de Android.	**No.**	Basada en la nube.	**Sí.**	Operación local.
Disponibilidad	**Sí.**	Operación local.	**No.**	Basada en la nube.	**Sí.**	Operación local.
Autenticidad	**Sí.**	Permisos granulares. Verificación del origen. Contemplado en el modelo de seguridad de Android.	**Sí.**	Mecanismo de fuente de datos.	**Sí.**	Acceso directo a los sensores.

Tabla I
COMPARATIVA DE SEGURIDAD ENTRE HEALTH CONNECT, GOOGLE FIT Y API DE SENSORES DE ANDROID.

- *Autenticidad:* La API de Sensores de Android garantiza la autenticidad de los datos al proporcionar acceso directo a los sensores del dispositivo. Esto implica que la información registrada proviene exclusivamente de la fuente correspondiente, sin riesgo de alteraciones en el origen.

Como se puede comprobar fácilmente en la tabla I, *Health Connect* es la única solución que cumple todos los pilares de seguridad. Esto nos permite confirmar que, efectivamente, *Health Connect* supone una mejora en términos de seguridad respecto a sus predecesoras.

V-B. Privacidad en Health Connect

La privacidad se basa en una serie de criterios cuya particularización al caso de las aplicaciones móviles la han llevado a cabo autoridades reguladoras, como la Agencia Española de Protección de Datos (AEPD) en España [24] o la Comisión Nacional de la Informática y las Libertades (CNIL) en Francia [25]. Para evaluar la privacidad en *Health Connect* y compararla con otras soluciones alternativas, se han seleccionado dos[11] criterios que, en nuestra opinión, son clave en el tratamiento de los datos de salud: la minimalidad de los datos y la capacidad del usuario para revocar su consentimiento.

Minimalidad de los datos Implica que solo se recojan aquellos estrictamente necesarios para la finalidad prevista, evitando un tratamiento excesivo; por ejemplo, una aplicación que realiza el seguimiento de los pasos diarios no debería acceder al sensor de presión arterial si solo cuenta el número de pasos.

Capacidad del usuario para revocar el consentimiento Garantiza que cualquier persona pueda retirar su autorización para el uso de sus datos en cualquier momento, por ejemplo, al revocar el acceso a alguno de sus datos de salud.

De manera similar a lo expuesto en la sección V-A, estudiar si es posible garantizar plenamente cada uno de estos criterios queda fuera del alcance de este estudio, ya que, en última instancia, es la implementación específica de la aplicación que gestiona los datos de salud la que determina su uso, compartición y almacenamiento. No obstante, evaluaremos si cada solución incorpora mecanismos que permitan a desarrolladores y usuarios cumplir con los criterios seleccionados.

Teniendo esta consideración en cuenta, y los criterios de privacidad seleccionados para realizar la comparación, en la tabla II se muestra una comparativa de privacidad entre *Health Connect*, Google Fit y la API de Sensores de Android.

Evaluación de los mecanismos ofrecidos por *Health Connect* para cumplir cada uno de los criterios seleccionados:

- *Minimalidad de los datos: Health Connect* ofrece un control granular sobre el acceso a los datos de salud, lo que constituye un mecanismo que facilita a desarrolladores y usuarios cumplir con el criterio de minimalidad de los datos.
- *Capacidad del usuario para revocar el consentimiento:* El control granular, a nivel de dato, que *Health Connect* proporciona a los usuarios y la operación local permite que dispongan de medios suficientes para ejercer sus derechos de revocación del consentimiento.

Evaluación de los mecanismos ofrecidos por Google Fit para cumplir cada uno de los criterios seleccionados:

- *Minimalidad de los datos:* Google Fit se basa en permisos de acceso generales, por lo que no proporciona mecanismos que faciliten cumplir con la minimalidad de los datos.
- *Capacidad del usuario para revocar el consentimiento:* Google Fit no proporciona mecanismos que faciliten a los usuarios revocar su consentimiento de manera sencilla. Aunque el usuario revoque los permisos de acceso en su dispositivo, no se garantiza que los datos ya almacenados en la plataforma dejen de ser utilizados o compartidos.

Evaluación de los mecanismos ofrecidos por la API de Sensores de Android para cumplir cada uno de los criterios seleccionados:

- *Minimalidad de los datos:* De modo similar a como sucede con Google Fit, la API de Sensores de Android se basa en un único permiso de acceso (`BODY_SENSORS`), por lo que no aporta mecanismos que faciliten cumplir con la minimalidad de los datos.
- *Capacidad del usuario para revocar el consentimiento:* Gracias al permiso de gestión de acceso a los sensores y

[11] Por razones de espacio se han seleccionado estos dos, críticos en nuestra opinión. Esto no significa que el resto de criterios recogidos en las dos referencias [24] y [25] no deban ser estudiados para disponer de un estudio más completo.

	Soluciones					
Criterio de Privacidad	***Health Connect***		**Google Fit**		**API de Sensores de Android**	
Minimalidad de los Datos	**Sí.**	Permisos granulares.	**No.**	Permiso general. Basada en la nube.	**No.**	Permisos general.
Capacidad de revocar el consentimiento.	**Sí.**	Permisos granulares. Operación local.	**No.**	Basada en la nube.	**Sí.**	Permiso general. Operación local.

Tabla II
COMPARATIVA EN PRIVACIDAD ENTRE HEALTH CONNECT, GOOGLE FIT Y API DE SENSORES DE ANDROID.

a la operación local de la API de Sensores de Android, los usuarios disponen de medios suficientes para ejercer sus derechos de revocación del consentimiento.

Como se puede comprobar fácilmente en la tabla II, *Health Connect* es la única solución que cumple todos los criterios de privacidad. El análisis de los mecanismos de privacidad utilizados por cada una de las soluciones estudiadas nos ha permitido confirmar que, efectivamente, *Health Connect* supone una mejora en términos de privacidad respecto a sus predecesoras.

VI. DISCUSIÓN

El análisis realizado se centra en las soluciones que utilizan las aplicaciones, aquello que se ejecuta en el contexto de un dispositivo. Un estudio completo sobre las garantías que una aplicación ofrece en términos de seguridad y privacidad requeriría un estudio de todos los elementos del ecosistema en que lleva a cabo sus funciones. En este tipo de estudio se considerarían las cualidades de las soluciones basadas en la nube, la comunicación con terceros, los mecanismos de recogida y transmisión de datos, etc. Esta perspectiva es más propia de una auditoría de seguridad o de una evaluación de impacto de privacidad de un tratamiento de datos personales. No es este el enfoque elegido en nuestro estudio, que ofrece una guía para quienes se preguntan qué implicaciones tiene en realidad *Health Connect* en general, independientemente de la aplicación o de cada tratamiento de datos de salud en concreto.

Como se indicó en la sección V-B, se han seleccionado dos criterios para evaluar la privacidad que se consideran críticos: minimalidad en el acceso a datos personales y capacidad de revocar el consentimiento. Tratándose de dispositivos móviles, son dos sobre los que los usuarios pueden tener información a través de su dispositivo (los permisos que conceden o revocan en la sección de Ajustes) y sobre los que pueden actuar. Esto lo hace especial, puesto que no requiere que recurran a figuras externas como el Delegado de Protección de Datos para comprobarlos (minimalidad) o ejercerlos (consentimiento), algo que no ocurre en otros entornos. Sin embargo, esto no impide que otros criterios, como la "licitud, transparencia y lealtad" y los restantes recogidos en las guías [24], [25] deban ser evaluados para tener un estudio completo.

El cifrado de datos en *Health Connect* plantea varios interrogantes sobre su seguridad y robustez. Como se vio en la sección IV, aunque la documentación oficial de la plataforma menciona que *Health Connect* emplea un sistema de cifrado para proteger la información, no se especifica qué algoritmos criptográficos se utilizan ni cómo se garantiza la confidencialidad de los datos. Esta falta de transparencia dificulta la evaluación de la seguridad del sistema, ya que un cifrado seguro debe ser público y auditable, lo que no es el caso en este contexto. En esa misma sección también se vio que es posible que *Health Connect* utilice el sistema de cifrado nativo de Android para proteger los datos, lo que aliviaría la necesidad de gestionar una clave de cifrado/descifrado de manera independiente, ya que Android implementa mecanismos de cifrado y gestión eficaces [6].

Sin embargo, el hecho de que no se especifique con detalle cómo *Health Connect* maneja estos procesos genera incertidumbre sobre su vulnerabilidad ante posibles ataques o brechas de seguridad. Aunque el cifrado proporcionado por Android puede ser eficaz, la falta de detalles específicos en la documentación de *Health Connect* sobre los mecanismos y algoritmos utilizados plantea serias dudas sobre la verdadera seguridad y privacidad de los datos de los usuarios.

Por último, conviene tener en cuenta que Google ha tenido un historial de retrasos en el cumplimiento de plazos y en la ejecución de cambios programados, como se evidenció en la adopción de la sección de seguridad de los datos en Google Play[12]. Por ello, no resulta sorprendente que se repita esta situación con el anuncio de que la API de Google Fit dejará de utilizarse. Dichos retrasos pueden repercutir negativamente en la adopción de *Health Connect*. Además, la falta de claridad en la comunicación y la incertidumbre sobre las fechas de implementación de cambios críticos en servicios tan fundamentales como los que gestionan los datos de salud pueden generar desconfianza en los usuarios y dificultar la planificación a largo plazo por parte de los desarrolladores.

VII. CONCLUSIONES

Este estudio ha evaluado la seguridad y privacidad de *Health Connect* en comparación con otras soluciones de gestión de datos de salud en Android, como Google Fit y la API de Sensores de Android. Se ha observado que *Health Connect* introduce mejoras significativas en términos de seguridad, especialmente en los pilares de confidencialidad, integridad y disponibilidad. Asimismo, se han identificado avances en términos de privacidad, particularmente en la minimización de los datos y la capacidad de revocar el consentimiento.

Sin embargo, la falta de claridad en la documentación de *Health Connect*, especialmente en algunos aspectos del cifrado, plantea incógnitas que requieren una mayor transparencia por parte de Google si se busca una mejor acogida de la plataforma por parte de usuarios y desarrolladores.

En resumen, *Health Connect* marca un paso importante en la gestión segura de los datos de salud en Android.

[12]https://support.google.com/googleplay/android-developer/answer/10787469

Sin embargo, es necesario mejorar la transparencia en su documentación.

VIII. Trabajo Futuro

Como trabajo futuro, se propone profundizar en la integración de *Health Connect* dentro de los estándares de seguridad para aplicaciones móviles, como OWASP MASVS v2.1.0 [26], con el objetivo de evaluar empíricamente su eficacia en la gestión segura de datos de salud. En particular, sería relevante analizar controles como `MASVS-STORAGE`, `MASVS-CRYPTO-2` y `MASVS-PRIVACY`, que abordan aspectos críticos de almacenamiento, cifrado y privacidad de la información. Este enfoque no solo contribuiría a fortalecer la protección de datos sensibles, sino que también establecería un marco de referencia para el desarrollo de aplicaciones móviles seguras, abriendo la puerta a investigaciones futuras sobre la escalabilidad de estos controles.

Adicionalmente, se propone ampliar el análisis de privacidad presentado en la sección V-B, incorporando un conjunto más amplio de criterios establecidos por organismos como la AEPD [24] y la CNIL [25]. Finalmente, se considera relevante incluir un análisis desde la perspectiva del usuario final, centrado en la usabilidad y comprensibilidad de la nueva interfaz de gestión de permisos granulares de *Health Connect*. Este tipo de estudio permitiría evaluar en qué medida el usuario medio es capaz de entender y gestionar adecuadamente sus configuraciones de privacidad, lo cual es crucial, ya que un sistema puede ser técnicamente seguro, pero ineficaz si no es comprensible para quienes lo utilizan.

Agradecimientos

Este trabajo se incluye en las actividades del Proyecto Estratégico de Ciberseguridad App-PI (*App Privacy Impact*): Un ecosistema para la evaluación del impacto de apps para dispositivos móviles sobre la privacidad y seguridad de sus usuarios, el cual se realiza al amparo de un convenio de colaboración entre la Universidad de Valladolid y la S.M.E. Instituto Nacional de Ciberseguridad de España M.P., S.A. para la promoción de proyectos estratégicos de Ciberseguridad en España, en el marco de los Fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation).

Referencias

[1] J. I. N. González and B. F. Morales, "Comunicación, Salud y Tecnología: mHealth," *Revista de Comunicación y Salud*, vol. 5, pp. 149–158, 2015. [Online]. Available: http://doi.org/10.35669/

[2] C. D. Hylender, P. Langlois, A. Pinto, and S. Widup, "Data Breach Investigations Report," Verizon, Tech. Rep., 2024.

[3] D. Seim, "El crecimiento explosivo de los dispositivos médicos portátiles," 2024, Último acceso: 10/02/2025. [Online]. Available: https://www.halcousa.com/es/El-crecimiento-explosivo-de-los-dispositivos-m%C3%A9dicos-port%C3%A1tiles./

[4] Backlinko, "iPhone vs. Android User & Revenue Statistics (2024)," 2024, Último acceso: 10/02/2025. [Online]. Available: https://backlinko.com/iphone-vs-android-statistics

[5] A. Aparicio, M. M. Martínez-González, and P. A. Criado-Lozano, "World App: El Secreto de tus Ojos está en los permisos Android," in *Actas de las IX Jornadas Nacionales de Investigación en Ciberseguridad, JNIC 2024*, A. M. R. Quintero, R. C. Guerrero, and Ángel J. Varela Vaca, Eds., 2024, pp. 245–252.

[6] R. Mayrhofer, J. V. Stoep, C. Brubaker, D. Hackborn, B. Bonné, G. S. Tuncay, R. P. Jover, and M. A. Specter, "The Android Platform Security Model (2023)," 2023. [Online]. Available: https://arxiv.org/abs/2503.02829

[7] G. Tuncay, "Android permissions: Evolution, attacks, and best practices," *IEEE Security & Privacy*, vol. PP, pp. 2–11, 2024.

[8] Parlamento Europeo y Consejo de la Unión Europea, "Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Reglamento General de Protección de Datos)," 2016, accedido el 18 de marzo de 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679

[9] U.S. Congress, "Health Insurance Portability and Accountability Act of 1996," 1996, accedido el 18 de marzo de 2025. [Online]. Available: https://www.hhs.gov/hipaa/

[10] C. Braghin, S. Cimato, and A. Della Libera, "Are mhealth apps secure? a case study," in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 02, 2018, pp. 335–340.

[11] M. Fan, L. Yu, S. Chen, H. Zhou, X. Luo, S. Li, Y. Liu, J. Liu, and T. Liu, "An Empirical Evaluation of GDPR Compliance Violations in Android mHealth Apps," in *2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE)*, 2020, pp. 253–264.

[12] J. Benjumea, J. Ropero, O. Rivera-Romero, E. Dorronzoro-Zubiete, and A. Carrasco, "Assessment of the fairness of privacy policies of mobile health apps: Scale development and evaluation in cancer apps," *JMIR Mhealth Uhealth*, vol. 8, no. 7, p. e17134, Jul 2020. [Online]. Available: http://mhealth.jmir.org/2020/7/e17134/

[13] A. Papageorgiou, M. Strigkos, E. Politou, E. Alepis, A. Solanas, and C. Patsakis, "Security and Privacy Analysis of Mobile Health Applications: The Alarming State of Practice," *IEEE Access*, vol. 6, pp. 9390–9403, 2018.

[14] G. Tangari, M. Ikram, I. W. B. Sentana, K. Ijaz, M. A. Kaafar, and S. Berkovsky, "Analyzing security issues of Android mobile health and medical applications," *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2074–2084, 08 2021. [Online]. Available: https://doi.org/10.1093/jamia/ocab131

[15] G. Tangari, M. Ikram, K. Ijaz, M. A. Kaafar, and S. Berkovsky, "Mobile health and privacy: cross sectional study," *BMJ*, vol. 373, 2021. [Online]. Available: https://www.bmj.com/content/373/bmj.n1248

[16] L. H. Iwaya, M. A. Babar, A. Rashid *et al.*, "On the privacy of mental health apps," *Empirical Software Engineering*, vol. 28, no. 2, 2023. [Online]. Available: https://doi.org/10.1007/s10664-022-10236-0

[17] K. Knorr and D. Aspinall, "Security testing for Android mHealth apps," in *2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 2015, pp. 1–8.

[18] J. J. Flors-Sidro, M. Househ, A. Abd-Alrazaq, J. Vidal-Alaball, L. Fernandez-Luque, and C. L. Sanchez-Bocanegra, "Analysis of Diabetes Apps to Assess Privacy-Related Permissions: Systematic Search of Apps," *JMIR Diabetes*, vol. 6, no. 1, p. e16146, 2021.

[19] Google, "Solicita acceso en segundo plano a los datos del sensor corporal," 2024, Último acceso: 20/02/2025. [Online]. Available: https://developer.android.com/health-and-fitness/guides/health-services/background-body-sensors?hl=es-419

[20] S. Miller, W. B. Glisson, M. Campbell, and S. M. Sittig, "Risk Analysis of Residual Protected Health Information of Android Telehealth Apps," in *AMCIS*, 2019.

[21] N. O. Ogwara, K. Petrova, M. L. B. Yang, and S. G. MacDonell, "A Risk Assessment Framework for Mobile Apps in Mobile Cloud Computing Environments," *Future Internet*, vol. 16, no. 8, p. 271, 2024.

[22] A. C. Mendoza De Los Santos, J. A. Rios Reyes, and R. Y. Vasquez Chiclayo, "Principales riesgos de seguridad en computación móvil en la nube: una revisión de la literatura," *Revista Científica de FAREM-Estelí: Medio ambiente, tecnología y desarrollo humano*, vol. Nº 45, pp. 206–220, 2023.

[23] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Tech. Rep. UCB/EECS-2009-28, Feb 2009. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[24] Agencia Española de Protección de Datos (AEPD), "El deber de informar y otras medidas de responsabilidad proactiva en apps para dispositivos móviles," Agencia Española de Protección de Datos, Tech. Rep., septiembre 2019. [Online]. Available: https://www.aepd.es/sites/default/files/2019-11/nota-tecnica-apps-moviles.pdf

[25] Commission Nationale Informatique & Libertés (CNIL), "Recommendation relative aux applications mobiles," Commission Nationale Informatique & Libertés, Tech. Rep., enero 2024. [Online]. Available: https://www.cnil.fr/sites/cnil/files/2024-09/recommandation-applications-mobiles.pdf

[26] OWASP Foundation, *OWASP Mobile Application Security Verification Standard (MASVS)*, 2024, versión 2.1.0. [Online]. Available: https://github.com/OWASP/owasp-masvs/releases/tag/v2.1.0

Aportes para el cumplimiento del Reglamento (UE) 2024/1689 en robótica y sistemas autónomos

Francisco J. Rodríguez Lera, Yoana Pita Lorenzo, David Sobrín Hidalgo, Laura Fernández Becerra
Universidad de León
Grupo de Robótica, EIIIA, Campus de Vegazana, S/N, 24071, León
fjrodl@unileon.es, {ypital00,inflfb00, dsobrh00}@estudiantes.unileon.es

Irene González Fernández, Jose Miguel Guerrero Hernández
Universidad Rey Juan Carlos,
Intelligent Robotics Lab, EIF, Fuenlabrada, 28943, Madrid, Spain
{irene.gonzalezf,josemiguel.guerrero}@urjc.es

***Resumen*—La ciberseguridad en robótica se erige como un aspecto clave dentro del Reglamento (UE) 2024/1689, también conocido como Ley de Inteligencia Artificial, que establece directrices específicas para sistemas inteligentes y automatizados. Una distinción fundamental en este marco regulatorio es la diferencia entre robots con Inteligencia Artificial (IA) y aquellos que operan mediante sistemas de automatización sin IA, dado que los primeros están sujetos a requisitos de seguridad más estrictos debido a su capacidad de aprendizaje y autonomía. En este trabajo, se analizan herramientas de ciberseguridad aplicables a sistemas robóticos avanzados, con especial énfasis en la protección de bases de conocimiento en arquitecturas cognitivas. Además, se propone un listado de herramientas básicas para garantizar la seguridad, integridad y resiliencia de estos sistemas, y se propone un caso práctico, enfocado al análisis de la gestión del conocimiento del robot, donde se definen diez criterios de evaluación que permitan asegurar la conformidad con el reglamento y reduzcan los riesgos en entornos de interacción humano-robot (HRI).**

***Index Terms*—Ciberseguridad, Herramientas, Inteligencia Artificial (IA), Reglamento, Robots**

Tipo de contribución: *Investigación original*

I. Introducción

La ciberseguridad en robótica es un aspecto que cobra especial importancia en el marco del Reglamento (UE) 2024/1689, conocido como *The AI Act* [1], el cual establece directrices específicas para sistemas automatizados e inteligentes. Una distinción fundamental dentro de este contexto es la diferencia entre robots que integran Inteligencia Artificial (IA) y aquellos que operan mediante sistemas de automatización sin IA. Esta diferenciación impacta directamente en los requisitos de seguridad y cumplimiento normativo, dado que los robots con IA están sujetos a regulaciones más estrictas en función de sus capacidades y nivel de autonomía.

La normativa europea de ciberseguridad adopta un enfoque basado en la evaluación del riesgo, lo que significa que las medidas de seguridad aplicadas a cada sistema dependen de su grado de exposición a amenazas. Por ejemplo, un robot que realiza acciones consideradas críticas, conducción autónoma o entretenimiento en escuelas de infantil, se enfrenta a riesgos significativamente mayores que un robot doméstico, lo que exige niveles de protección diferenciados.

Este enfoque flexible permite equilibrar la innovación con la seguridad, evitando la imposición de requisitos innecesarios para sistemas de bajo riesgo, al tiempo que se refuerzan las protecciones en entornos críticos. En este contexto, el uso de herramientas de auditoría de ciberseguridad es clave para garantizar la seguridad de los sistemas robóticos y su cumplimiento con las normativas vigentes.

Este trabajo analiza los riesgos de ciberseguridad que afectan a los robots autónomos en el contexto del Reglamento (UE) 2024/1689, identificando las amenazas específicas que surgen de la regulación existente y su impacto en la robótica cognitiva. Considerando arquitectura cognitiva como la estructura funcional que organiza los distintos módulos encargados de dotar al robot de capacidades como la percepción, el razonamiento, la planificación y el aprendizaje, permitiendo su adaptación al entorno y la toma de decisiones autónoma.

La metodología propuesta permite evaluar aspectos críticos de seguridad, incluyendo identificación de riesgos, protección de datos sensibles, gestión de actualizaciones y parches, resiliencia ante ciberataques, cumplimiento normativo, gestión de vulnerabilidades, planificación de respuesta a incidentes y auditoría continua. Este enfoque facilita la integración de medidas de seguridad en sistemas autónomos, garantizando su conformidad con las normativas europeas y fortaleciendo su capacidad de operar en entornos de interacción humano-robot.

A través de la integración de estos mecanismos de seguridad, se garantiza la integridad y confiabilidad de la toma de decisiones en sistemas robóticos, evitando modificaciones no autorizadas en la base de conocimiento y fortaleciendo la resiliencia de la Interacción Humano-Robot (HRI) ante amenazas cibernéticas. Este trabajo proporciona un marco de referencia para el desarrollo de sistemas cognitivos robóticos seguros y confiables.

El artículo se estructura del siguiente modo: la sección II revisa el estado del arte en ciberseguridad robótica, con énfasis en sistemas como ROS 2, y la protección de su arquitectura cognitiva. La sección III analiza las directrices del Reglamento para clasificar sistemas robóticos según su nivel de riesgo y establecer requisitos. En IV, se adopta un enfoque multinivel para abordar la seguridad en robots con IA, integrando aspectos éticos y regulatorios. En V, se examina una arquitectura híbrida en robots autónomos y sus desafíos de seguridad. La sección VI define los criterios de evaluación para el análisis de la base de conocimiento para

el cumplimiento del reglamento 2024/1689. Finalmente, la sección VII enumera las herramientas técnicas específicas para la revisión y análisis de los criterio y VIII recoge las principales aportaciones y plantea líneas futuras.

II. Estado del Arte

Estudios como los planteados en [2], [3] han establecido bases fundamentales en la seguridad en infraestructuras críticas, en particular a través de marcos de ciberseguridad como el desarrollado por el Instituto Nacional de Estándares y Tecnología (NIST). Este marco proporciona un enfoque estructurado para evaluar, mitigar y gestionar los riesgos de ciberseguridad en sistemas altamente conectados y tecnológicamente avanzados. A través de modelos como el *Cybersecurity Framework*, se establecen mecanismos flexibles y adaptativos para la identificación, protección, detección, respuesta y recuperación ante ciberamenazas en infraestructuras clave, incluyendo la robótica. Nuestra investigación se apoya en estos marcos para extender su aplicación a los sistemas de Inteligencia Artificial desplegados en un robot y en particular a la seguridad de la base de conocimiento del robot, considerando la complejidad de los sistemas y modelos de IA utilizados en este contexto.

En este contexto, se podría considerar el trabajo de Victor et al. [4] como uno de los motores de esta investigación. En el se analiza la seguridad en el contexto de Robot Operating System (ROS) y su evolución hacia ROS 2. Aunque proporciona una visión técnica sólida, su tratamiento de la inteligencia artificial en el robot es limitado. Nuestra contribución en este sentido es ampliar el modelo propuesto, integrando la seguridad de la base de conocimiento en los sistemas deliberativos y subsimbólicos del robot.

Por otro lado, estudios como [5] abordan la necesidad de construir sistemas robóticos con IA desplegada confiables, centrándose en ciberseguridad, ética y regulación. Sin embargo, algunas de estas aproximaciones tienden a enfocarse exclusivamente en el aspecto ético, sin profundizar en los desafíos técnicos. En nuestro trabajo, abordamos directamente la dimensión técnica, proporcionando una evaluación estructurada de las vulnerabilidades y medidas de protección para la base de conocimiento del robot.

Desde una perspectiva más cercana a la toma de decisiones, [6] plantea un modelo que profundiza en este aspecto, aunque no enfatiza la importancia del conocimiento previo que posee el robot. Nuestro trabajo complementa esta aproximación, asegurando que la seguridad de la base de conocimiento se integre de manera efectiva en la arquitectura cognitiva del robot.

Por otra parte, [7] presenta el SecuRoPS Framework, un modelo arquitectónico detallado para la integración de robots en entornos públicos. Aunque este marco incluye una perspectiva de seguridad amplia, nuestra contribución específica se centra en la protección del conocimiento del robot, asegurando la integridad y resiliencia de los datos almacenados y procesados tanto en la infraestructura local como en la nube.

III. Aplicación del Reglamento (UE) 2024/1689 a la Robótica con IA

El Reglamento (UE) 2024/1689 [8], [9] establece directrices específicas para la seguridad y confiabilidad de los sistemas autónomos e inteligentes. En el contexto de la interacción humano-robot (HRI), los requisitos clave incluyen la clasificación del riesgo del sistema de IA, la implementación de medidas de ciberseguridad, y el cumplimiento de estándares de supervisión y transparencia.

El reglamento clasifica los sistemas robóticos con IA desplegada en diferentes niveles de riesgo, determinando los requisitos que deben cumplir:

- Riesgo inaceptable: Sistemas robóticos con IA desplegada que representan una amenaza directa para la seguridad, los derechos fundamentales o el bienestar de las personas. Estos sistemas están prohibidos e incluyen:
  - Manipulación y engaño perjudiciales basados en IA.
  - Explotación de vulnerabilidades con fines perjudiciales.
  - Puntuación social y predicción del riesgo de infracción penal.
  - Raspado masivo de datos biométricos para bases de reconocimiento facial.
  - Identificación biométrica remota en tiempo real para fines policiales en espacios públicos.
  - Reconocimiento de emociones en lugares de trabajo e instituciones educativas.
  - Categorización biométrica para inferir características protegidas.
- Sistemas de alto riesgo: Robots con IA que pueden afectar la seguridad física o los derechos fundamentales de las personas. Ejemplos incluyen:
  - Componentes de IA en infraestructuras críticas (como vehículos autónomos).
  - IA en robots desplegados en instituciones educativas.
  - Componentes de seguridad en productos médicos y cirugía asistida por robot.
  - IA aplicada a la identificación biométrica y el reconocimiento de emociones.
- Sistemas de riesgo limitado: Requieren transparencia en su funcionamiento, pero no presentan un riesgo crítico.
- Sistemas de riesgo mínimo: Pueden operar con menor supervisión, siempre que no interfieran en la seguridad o privacidad de los usuarios. Los robots de entretenimiento, sin comportamientos asociados a la personalización, podrían entrar aquí.

Si el robot autónomo se clasifica como de alto riesgo, el cumplimiento del reglamento exige la implementación de medidas estrictas:

- Gestión de riesgos: Identificación y mitigación de vulnerabilidades potenciales en el sistema robótico con IA desplegada.
- Documentación técnica: Mantenimiento de registros detallados sobre el desarrollo, operación y seguridad del sistema.
- Registro de datos: Almacenamiento de eventos relevantes durante el funcionamiento del robot para auditoría y trazabilidad.
- Supervisión humana: Mecanismos que permitan intervención humana en caso de fallos o riesgos de seguridad.
- Transparencia y explicabilidad: Información clara sobre las capacidades, limitaciones y decisiones del sistema autónomo.

- Robustez y seguridad: Protección ante errores, ataques adversariales y manipulación de datos.

Además de la regulación sobre la IA, el reglamento exige garantizar la seguridad física y la protección contra ciberataques:

- Seguridad del hardware: Implementación de controles físicos, como cerraduras y monitoreo de acceso al robot.
- Protección contra software malicioso: Supervisión continua y detección de amenazas en la infraestructura del robot.
- Gestión de identidad y acceso: Autenticación robusta para restringir el acceso a los sistemas internos.
- Seguridad de redes: Uso de VPNs y cifrado para proteger la comunicación entre el robot y la nube.
- Protección de datos: Implementación de mecanismos de cifrado y controles de acceso para la información almacenada.
- Auditorías y supervisión: Implementación de sistemas SIEM y análisis periódicos de vulnerabilidades.

Si el robot utiliza servicios en la nube para procesamiento o almacenamiento de datos, el reglamento exige:

- Seguridad en la nube: Implementación de cifrado y acceso restringido a datos sensibles.
- Monitoreo de acceso: Registro de accesos y detección de anomalías en la gestión del conocimiento.
- Auditoría de conformidad: Evaluación periódica para garantizar la alineación con los requisitos normativos.

El cumplimiento de estas normativas exige un monitoreo y verificación constantes de la seguridad en los sistemas autónomos. Para ello, las auditorías de ciberseguridad juegan un papel fundamental, ya que permiten detectar vulnerabilidades, evaluar la resiliencia de los sistemas ante ataques y garantizar el cumplimiento de los requisitos establecidos en el Reglamento (UE) 2024/1689.

IV. Riesgos en la IA de los Robots

La seguridad en sistemas robóticos con IA debe abordarse desde múltiples capas, siguiendo una estructura similar al Robot Security Framework [4]. A continuación, se presentan los principales aspectos a considerar en cada capa y vamos a contribuir teniendo en cuenta la aplicación de consideraciones éticas y regulatorias.

IV-A. Capa Física

Aunque la IA es predominantemente un componente de software, su integración puede afectar la seguridad física del robot. Se deben considerar:

- **Protección del hardware**: Protección de hardware especializado (TPU, GPU) contra manipulaciones.
- **Resguardo de sensores**: Evitar sabotajes o interferencias físicas en sensores utilizados por la IA.
- **Redundancia en percepción**: Implementación de múltiples sensores para mejorar la resiliencia a fallos.

IV-B. Capa de Red

El uso de IA en robots implica la transmisión de datos de entrenamiento e inferencias a través de redes, lo que introduce riesgos de seguridad:

- **Cifrado de modelos y datos**: Protección de modelos y datos de entrenamiento en tránsito y en reposo.
- **Defensa contra ataques adversariales en la red**: Prevención de ataques de envenenamiento de datos y manipulación de inferencias.
- **Autenticación de dispositivos**: Implementación de autenticación fuerte para evitar accesos no autorizados.

IV-C. Capa de Firmware

La IA en robots a menudo se ejecuta en hardware embebido con firmware especializado:

- **Ejecución en entornos confiables (TEE)**: Uso de entornos de ejecución seguros para evitar manipulación de modelos.
- **Actualización segura de modelos**: Verificación de integridad y autenticidad en actualizaciones de firmware y modelos de IA.

IV-D. Capa de Aplicación

La IA impacta significativamente la seguridad en la capa de aplicación, donde se gestionan los modelos, inferencias y decisiones del robot:

- **Explicabilidad y Transparencia**: Implementación de técnicas como LIME o SHAP para auditar decisiones de IA.
- **Defensas contra ataques adversariales**: Entrenamiento robusto y validaciones de entrada para evitar manipulación de inferencias.
- **Monitoreo de comportamiento**: Implementación de detección de drift y auditoría de modelos en producción.
- **Privacidad y protección de datos**: Cumplimiento con GDPR y otros marcos regulatorios.
- **Control de acceso a modelos**: Restricción del acceso a modelos y API de inferencias mediante autenticación robusta.
- **Cifrado en comunicaciones**: Protección de inferencias transmitidas a la nube o a otros dispositivos.

IV-E. Consideraciones Éticas y Regulatorias

Además de la seguridad técnica, la integración de IA en robots plantea cuestiones éticas y legales:

- **Cumplimiento con estándares internacionales**: Asegurar conformidad con normativas como NIST AI RMF, ISO/IEC 23894.
- **Responsabilidad y trazabilidad**: Registro de decisiones críticas y análisis de fallos.
- **Implementación de mecanismos de seguridad**: Implementación de *kill switch* en caso de mal funcionamiento.

V. Consideraciones de seguridad en arquitecturas cognitivas

Las arquitecturas cognitivas en robótica se organizan en niveles funcionales que dotan al robot de percepción avanzada, toma de decisiones y adaptación. El sistema deliberativo planifica a largo plazo usando una memoria simbólica para representar el conocimiento. El sistema de comportamientos ejecuta acciones reactivas en tiempo real, apoyado en una memoria subsimbólica basada en la percepción. Por otro lado, muchos de estos sistemas tienen componentes de ejecución en

la nube, permitiendo acceder a modelos y datos externos, facilitando la actualización y mejora continua de las capacidades cognitivas.

La implementación de una arquitectura híbrida en robots autónomos [10], [11], que combine procesamiento local con servicios en la nube, plantea desafíos significativos en el contexto del Reglamento (UE) 2024/1689 sobre inteligencia artificial. A continuación, se destacan las principales problemáticas:

- Clasificación de Riesgo: El reglamento establece una clasificación de los sistemas robóticos con IA desplegada basada en el nivel de riesgo que representan para los derechos fundamentales y la seguridad. Los sistemas de alto riesgo están sujetos a requisitos más estrictos. Las arquitecturas híbridas deben evaluarse para determinar su nivel de riesgo y cumplir con las obligaciones correspondientes.
- Transparencia y Explicabilidad: El reglamento exige que los sistemas robóticos con IA desplegada sean transparentes y explicables, permitiendo a los usuarios comprender cómo se toman las decisiones. En una arquitectura híbrida, donde parte del procesamiento ocurre en la nube, garantizar esta transparencia puede ser complejo debido a la posible opacidad de los procesos en la nube.
- Supervisión Humana: Es fundamental que los sistemas de IA permitan la supervisión humana efectiva. En arquitecturas híbridas, la supervisión puede verse dificultada por la distribución de componentes entre el dispositivo local y la nube, lo que podría afectar la capacidad de intervención humana en tiempo real.
- Ciberseguridad: La conexión a la nube introduce riesgos adicionales de ciberseguridad. El reglamento requiere que los sistemas robóticos con IA desplegada implementen medidas para protegerse contra ataques que puedan comprometer su funcionamiento o la seguridad de los datos. Las arquitecturas híbridas deben abordar estos riesgos de manera integral.
- Gestión de Datos y Privacidad: El reglamento enfatiza la protección de datos y la privacidad. En una arquitectura híbrida, los datos pueden transferirse entre el robot y la nube, lo que plantea desafíos en términos de control y protección de la información personal.
- Responsabilidad y Rendición de Cuentas: Determinar la responsabilidad en caso de fallos o daños es más complejo en sistemas híbridos debido a la participación de múltiples actores (desarrolladores, proveedores de servicios en la nube, operadores, etc.). El reglamento exige una clara asignación de responsabilidades.

En este artículo, nos enfocaremos en la gestión del conocimiento del robot, específicamente en los mecanismos de información asociados a lo componentes de memoria y la infraestructura de datos en la nube, tal como se representa en la figura. Para ello, utilizaremos el enfoque presentado en este estudio [4].

VI. Criterios de Evaluación y Análisis de Amenazas

Para este análisis, consideraremos la gestión del conocimiento en un robot cognitivo basado en ROS 2, centrándonos en los mecanismos de almacenamiento y acceso a la información tanto en la memoria local como en la infraestructura en la nube. En este contexto, el robot utiliza una Base de Conocimiento (*Knowledge Base*), que se asocia a las memorias presentadas en la figura anterior y que integra múltiples fuentes de conocimiento:

- **Memoria simbólica**: Contiene representaciones explícitas, legibles y manipulables mediante reglas lógicas o lenguajes formales. Se emplea principalmente en la planificación deliberativa y la toma de decisiones de alto nivel. Un ejemplo típico es el uso de lenguajes como *PDDL*, que permiten definir acciones, objetivos y restricciones del entorno de forma estructurada.
- **Memoria subsimbólica**: Gestiona representaciones implícitas, codificadas mediante parámetros de modelos entrenados (por ejemplo, redes neuronales profundas). Esta forma de conocimiento es esencial para tareas perceptivas, motoras y de aprendizaje adaptativo. Incluye:
  - **Modelos de visión por computador**: como `YOLO_ROS` [12], utilizados para la detección de objetos en tiempo real a partir de imágenes de cámara, integrando modelos de Ultralytics en sistemas robóticos.
  - **Modelos de audio y voz**: por ejemplo, `Whisper_ROS` [13], utilizado para reconocimiento de voz en entornos ruidosos o multimodales.

Estos son los sistemas de memoria clásicos, a los que tenemos que añadir también lo que se podría considerar como Memoria híbrida. Corresponde a sistemas que, aunque están implementados con estructuras subsimbólicas, operan funcionalmente sobre contenido simbólico. Este es el caso de los modelos de lenguaje de gran escala (*LLMs*), como los utilizados en `Llama_ROS` [14], que permiten generar, completar o interpretar instrucciones en lenguaje natural. Estos modelos actúan como interfaces simbólicas, pero con una base computacional subsimbólica (p.e. embeddings) lo que les confiere un carácter mixto que requiere una atención especial en cuanto a seguridad, explicabilidad y trazabilidad del conocimiento generado.

Estas memorias, y sus modelos asociados, presentan múltiples vulnerabilidades que pueden comprometer la seguridad del sistema. Existen ataques dirigidos tanto a modelos simbólicos como subsimbólicos, afectando la toma de decisiones y la fiabilidad del conocimiento del robot. Según aproximaciones como las propuestas por OWASP [15], [16]. Los principales riesgos incluyen la manipulación de datos de entrada para alterar el comportamiento del modelo, el envenenamiento de datos en el entrenamiento y la inferencia de información sensible a partir de modelos entrenados. Además, la alteración de salidas y la explotación de vulnerabilidades en la cadena de toma de decisiones representan amenazas significativas que pueden comprometer la seguridad del conocimiento almacenado y de las decisiones tomadas.

Para mitigar estos riesgos, es necesario adoptar los siguientes criterios para la evaluación y protección de cada componente con acceso y o mantenimiento al conocimiento del robot.

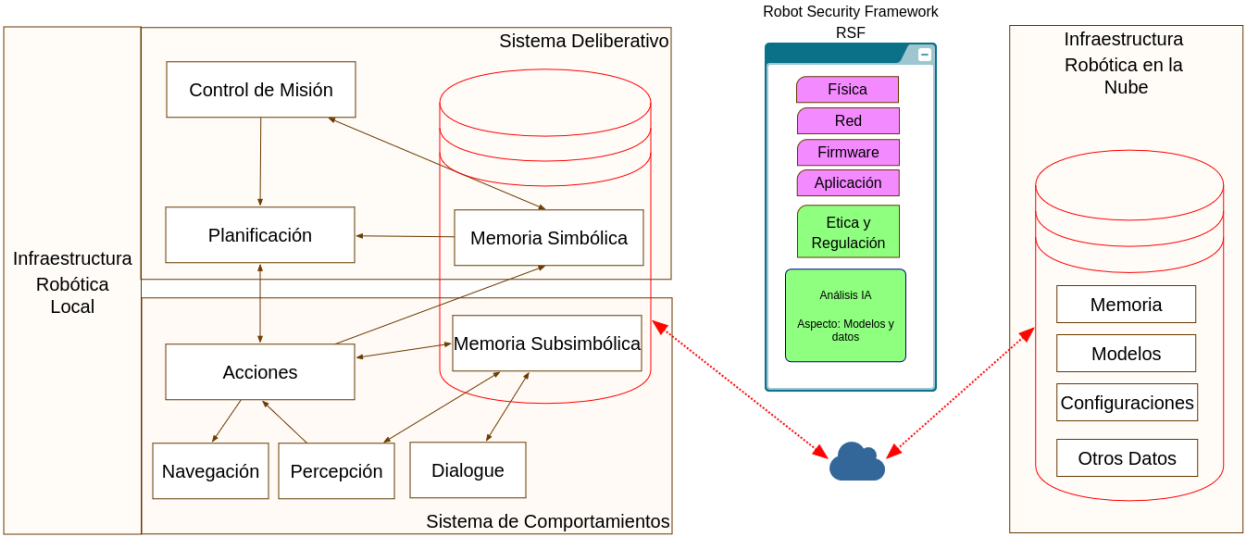


Figura 1. Ejemplo de arquitectura cognitiva para robots con análisis RSF y con nuestra contribución para el análisis del conocimiento.

VI-A. Criterio 1: Memoria Simbólica

Objetivo: Evaluar la estructura y coherencia de representaciones simbólicas en planificación y razonamiento lógico.

Justificación: Los modelos simbólicos como PDDL o las ontologías permiten la planificación basada en reglas explícitas y estructuradas. Tanto el dominio como el problema, ficheros básicos de funcionamiento de PDDL, puede ser modificado en tiempo de ejecución, dando lugar a problemas de expresividad y explicabilidad o planificaciones erróneas.

Método:

1. Analizar la ontología utilizada (clases, relaciones, reglas).
2. Evaluar la consistencia lógica del modelo mediante pruebas de inferencia.
3. Medir la eficiencia del planificador en diferentes escenarios.
4. Examinar la integración del conocimiento simbólico con representaciones subsimbólicas.
5. Identificar riesgos de manipulación de reglas en PDDL que puedan afectar la toma de decisiones.

VI-B. Criterio 2: Memoria Subsimbólica

Objetivo: Analizar la fiabilidad y precisión de los datos de configuración y parámetros almacenados.

Justificación: Los robots dependen de archivos de configuración para parametrizar su comportamiento. Datos mal calibrados o corruptos pueden comprometer su rendimiento y seguridad. Por ejemplo, ficheros de configuración de Nav2

Método:

1. Verificar la coherencia y formato de los ficheros de configuración.
2. Evaluar la sensibilidad del sistema a errores en los parámetros.
3. Medir el impacto de variaciones en los datos de entrada sobre el comportamiento del robot.
4. Implementar estrategias de detección de corrupción de datos.
5. Aplicar técnicas de cifrado y control de acceso en bases de conocimiento.

VI-C. Criterio 3: Sistemas de Aprendizaje

Objetivo: Examinar la robustez, precisión y posibles vulnerabilidades en modelos entrenados para percepción.

Justificación: Los modelos de Machine Learning son esenciales para la percepción del entorno en robots. Sin embargo, son susceptibles a ataques adversariales, sesgo en los datos y problemas de interpretabilidad.

Método:

1. Evaluar la precisión del modelo en distintos escenarios.
2. Identificar sesgos en los datos de entrenamiento.
3. Analizar la resiliencia ante ataques adversariales.
4. Medir el impacto del modelo en la toma de decisiones del robot.
5. Implementar firmas digitales para garantizar la integridad de los modelos, por ejemplo los de YOLO.
6. Aplicar técnicas de aprendizaje adversarial para fortalecer modelos de visión.

VI-D. Criterio 4: IA Generativa

Objetivo: Evaluar la capacidad de generalización, interpretación y susceptibilidad a ataques adversariales en modelos basados en representaciones densas y aprendizaje profundo. De este modo se cubre las nuevas opciones de Embeddings y Large Language Models.

Justificación: Los modelos de lenguaje y embeddings mejoran la capacidad del robot para interpretar información no estructurada. Sin embargo, pueden ser manipulados mediante inyección de instrucciones, presentar sesgos y carecer de interpretabilidad.

Método:

1. Evaluar la coherencia y precisión de las respuestas generadas.
2. Identificar vulnerabilidades ante ataques de inyección de instrucciones.
3. Medir la capacidad de adaptación a diferentes dominios.
4. Analizar el impacto del modelo en la autonomía del robot.
5. Aplicar firmas digitales para garantizar la integridad de modelos de LLMs.
6. Implementar mecanismos de fallback para generar respuestas seguras en LLMs.

VI-E. Criterio 5: Seguridad y Robustez en el Almacenamiento y Recuperación del Conocimiento

Objetivo: Evaluar la protección del conocimiento ante corrupción, accesos no autorizados y pérdidas de información.

Justificación: El conocimiento almacenado en robots puede ser un objetivo de ataques o verse comprometido por fallos técnicos. La seguridad y robustez del almacenamiento son esenciales para garantizar la fiabilidad operativa.

Método:

1. Analizar los mecanismos de control de acceso y autenticación.
2. Evaluar estrategias de recuperación ante pérdida de datos.
3. Medir la resiliencia frente a corrupción de información y ataques externos.
4. Aplicar medidas de sanitización en datos de entrenamiento de LLMs.
5. Implementar auditorías para garantizar el cumplimiento de normativas como GDPR y NIST.

VI-F. Criterio 6: Evaluación en Entornos Reales

Objetivo: Validar el desempeño del conocimiento en escenarios operativos.

Justificación: Un sistema cognitivo puede comportarse bien en simulaciones pero fallar en entornos reales debido a variaciones no previstas. Evaluar el desempeño en el mundo real es crucial.

Método:

1. Comparar el rendimiento del sistema en simulaciones y pruebas reales.
2. Evaluar la adaptabilidad ante condiciones adversas.
3. Medir el impacto de la calidad del conocimiento en la eficacia operativa del robot.
4. Implementar detección de ataques adversariales en modelos de visión.
5. Realizar pruebas de penetración y simulaciones de ataques en entornos de desarrollo.
6. Establecer mecanismos de rollback para restaurar versiones seguras de los modelos.

VI-G. Criterio 7: Auditorías y Supervisión en Infraestructura Local

Objetivo: Garantizar el monitoreo continuo y la validación de eventos de seguridad en la infraestructura local del robot.

Justificación: Es necesario supervisar el conocimiento almacenado y los procesos que acceden a él para detectar anomalías y vulnerabilidades que puedan comprometer la seguridad operativa del robot.

Método:

1. Integrar SIEM (Wazuh) para monitoreo continuo de eventos en la Knowledge Base local.
2. Definir métricas de seguridad para evaluar la estabilidad del sistema en el robot.
3. Monitorear las actualizaciones manuales o automáticas asociadas a los modelos y datos: desarrolladores, modelos de fuentes públicas (HuggingFace).
4. Implementar auditorías periódicas en los sistemas de almacenamiento y procesamiento de conocimiento.

VI-H. Criterio 8: Auditorías y Supervisión en Infraestructura en la Nube

Objetivo: Supervisar el acceso, integridad y seguridad del conocimiento almacenado en la infraestructura en la nube.

Justificación: Los sistemas en la nube presentan riesgos adicionales debido a la exposición a ataques externos, la dependencia de servicios de terceros y la necesidad de garantizar la integridad del conocimiento.

Método:

1. Monitorear el acceso y la integridad de modelos y datos almacenados en la nube.
2. Implementar auditorías periódicas para validar la integridad del conocimiento en entornos cloud.
3. Monitorear las actualizaciones manuales o automáticas asociadas a los modelos y datos: desarrolladores, modelos de fuentes públicas (HuggingFace).
4. Evaluar el cumplimiento de normativas de ciberseguridad en la gestión del conocimiento en la nube.

VI-I. Criterio 9: Gestión de Vulnerabilidades

Objetivo: Identificar, mitigar y gestionar vulnerabilidades en el conocimiento y los modelos utilizados por el robot.

Justificación: Las vulnerabilidades en la base de conocimiento pueden ser explotadas para manipular el comportamiento del robot, comprometer la seguridad del sistema y afectar la toma de decisiones.

Método:

1. Integrar herramientas de escaneo de vulnerabilidades en ROS 2 y su Sistema Operativo.
2. Implementar listas blancas y control de acceso a los sistemas de gestión del conocimiento del robot (cualquiera de las memorias).
3. Validar parches de seguridad antes de su implementación.
4. Establecer planes de respuesta ante vulnerabilidades detectadas.

VI-J. Criterio 10: Protección de Datos Sensibles y Biometría

Objetivo: Garantizar la seguridad, privacidad y cumplimiento normativo en el almacenamiento, procesamiento y acceso a datos biométricos y otra información sensible dentro de la base de conocimiento del robot.

Justificación El almacenamiento de datos biométricos dentro del robot y en la nube introduce riesgos de privacidad, posible uso indebido y requerimientos de cumplimiento normativo, especialmente con el Reglamento (UE) 2024/1689 y normativas como el GDPR [17]. La exposición de estos datos puede generar vulnerabilidades críticas, por lo que su

gestión debe garantizar cifrado, acceso restringido y auditoría continua.

Método:

1. Cifrado de datos en reposo y en tránsito: Aplicar técnicas de cifrado de extremo a extremo en los datos biométricos almacenados en la base de conocimiento y transmitidos a la nube (TLS 1.3).
2. Gestión de identidad y control de acceso: Implementar autenticación multifactor (MFA) y roles de acceso estrictos para evitar accesos no autorizados a los datos biométricos.
3. Monitorización y auditoría: Integrar sistemas de detección de accesos no autorizados (SIEM) y auditorías periódicas para detectar posibles filtraciones de datos biométricos.
4. Minimización y anonimización de datos: Aplicar principios de minimización de datos, evitando el almacenamiento innecesario de información biométrica y, cuando sea posible, anonimizando los datos.
5. Cumplimiento normativo: Verificar que los procedimientos de gestión de datos biométricos cumplan con el Reglamento (UE) 2024/1689, el GDPR y otras regulaciones de privacidad aplicables.
6. Seguridad en la nube: Asegurar que cualquier dato biométrico recibido y almacenado en la nube siga estándares de protección como cifrado avanzado (AES256) y acceso restringido utilizando IAM (Identity and Access Management).

VII. Aplicación Técnica del Marco de Ciberseguridad en Robótica con IA

Este estudio establece un primer marco de referencia integral para la ciberseguridad en sistemas robóticos con componentes IA desplegados, alineado con el Reglamento (UE) 2024/1689. A lo largo del artículo se han abordado aspectos críticos como la identificación de riesgos, la protección de datos, la gestión de actualizaciones, la resiliencia ante ciberataques y el cumplimiento de estándares de seguridad. En este contexto, y tras definir los criterios de evaluación y análisis aplicables a los sistemas de conocimiento de una arquitectura cognitiva híbrida, se presentan diversas soluciones técnicas orientadas a proteger dichos sistemas.

Algunas de las herramientas de auditoría de ciberseguridad específicas para estos sistemas incluyen:

- **Sistemas de Detección de Intrusiones (IDS) y Prevención (IPS)**: Monitorean el tráfico de red en busca de actividades sospechosas y pueden bloquear ataques en tiempo real [18], [19].
- **Firewall de Próxima Generación (NGFW)**: Proporcionan una capa adicional de seguridad al inspeccionar el tráfico de red y aplicar políticas de seguridad avanzadas [20] [21].
- **Soluciones de Gestión de Información y Eventos de Seguridad (SIEM)**: Recopilan y analizan datos de seguridad de múltiples fuentes para detectar y responder a incidentes de seguridad [22].
- **Autenticación Fuerte y Control de Acceso**: Implementar autenticación multifactor y políticas de control de acceso estrictas para proteger los sistemas críticos. *Ejemplo*: RSA SecurID [23], Okta [24].
- **Segmentación de Redes**: Dividir la red en segmentos más pequeños para limitar el alcance de un posible ataque. *Ejemplo*: Cisco TrustSec [25], VMware NSX [26].
- **Sistemas de Gestión de Parches**: Herramientas que ayudan a mantener los sistemas actualizados con los últimos parches de seguridad. *Ejemplo*: Microsoft System Center Configuration Manager (SCCM) [27], Ivanti Patch Management [28].
- **Monitorización y Detección de Amenazas**: Soluciones que permiten la monitorización continua de los sistemas para detectar y responder a amenazas en tiempo real. *Ejemplo*: Darktrace [29], CrowdStrike Falcon [30].
- **Soluciones de Respuesta y Recuperación ante Incidentes**: Herramientas que ayudan a gestionar y mitigar los efectos de un incidente de seguridad. *Ejemplo*: CyberArk [31], Trellix(anteriormente FireEye y McAfee Enterprise) [32].
- **Cifrado y control de acceso en bases de conocimiento**: Herramientas como **SROS2** para asegurar que solo los nodos autorizados accedan a esos datos, controlando el acceso al servicio o nodo que los maneja; cifrados clásicos sobre discos usando LUKS; de BBDD utilizando SQLlite; cifrados GPG sobre ficheros; uso de SealFs [33] para garantizar que los datos no han sido manipulados o sistemas basados en BlockChain para salvaguardar el estado a lo largo del tiempo en sistemas en la nube.

Estas herramientas son el medio para implementar medidas de protección ajustadas a los riesgos de los sistemas robóticos con IA y a los marcos normativos aplicables. Ellas vinculan el análisis teórico de riesgos, vulnerabilidades y cumplimiento con soluciones prácticas que refuerzan la seguridad y la integridad operativa. Al integrarlas, garantizamos un enfoque de seguridad holístico —desde lo conceptual hasta lo operativo— en entornos críticos y altamente automatizados.

VIII. Conclusiones

Este estudio establece un marco de referencia integral para la ciberseguridad en sistemas robóticos con IA desplegada, alineado con el Reglamento (UE) 2024/1689. En este artículo se han abordado aspectos críticos como la identificación de riesgos, la protección de datos, la gestión de actualizaciones, la resiliencia ante ciberataques y el cumplimiento de estándares de seguridad. Históricamente, herramientas como Secure Robot Operating System 2 (SROS2), han sido clave para garantizar comunicaciones cifradas, control de acceso robusto y verificación de integridad mediante firmas digitales. Sin embargo, en el nuevo entorno es necesario integrar estrategias que se adapten a las nuevas áreas de enfoque de la Inteligencia Artificial como la IA generativa, aprendizaje automático, etc, que permitan detectar ataques adversariales en visión artificial, monitorizar amenazas por anomalías y aplicar buenas prácticas en los robots que las despliegan.

Estas medidas refuerzan la integridad y confiabilidad de los sistemas robóticos, evitando accesos no autorizados y fortaleciendo su resiliencia frente a amenazas cibernéticas. Además, promueven un equilibrio entre la innovación tecnológica y el cumplimiento normativo, consolidando un enfoque responsable para el desarrollo de sistemas seguros y adaptables.

En este contexto, se propone un conjunto de criterios de evaluación para la validación empírica que permita confirmar la efectividad de las herramientas, en gran parte heredado de otros entornos de ciberseguridad, y queda abierto al análisis pormenorizado de las estrategias a adoptar, siendo necesario un estudio más profundo para cada uno de los criterios generales aquí propuestos.

Agradecimientos

Proyecto DMARCE - PID2021-126592OB-C21, PID2021-126592OB-C22 financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE. Plan de Recuperacion, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation) gracias al proyecto TESCAC (Trazabilidad y Explicabilidad en Sistemas Autonomos para la mejora de la Ciberseguridad) concedido por INCIBE a la Universidad de León. David Sobrín Hidalgo agradece a la Universidad de León la beca proporcionada para financiar sus estudios de doctorado. Proyecto CORESENSE, con financiación del Programa de Investigación e Innovación Horizonte Europa de la Unión Europea (Acuerdo de Subvención n.º 101070254).

Referencias

[1] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 on Artificial Intelligence Act," 2024, Último acceso: 19 de marzo de 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[2] A. Sedgewick, "Framework for improving critical infrastructure cybersecurity, version 1.0," National Institute of Standards and Technology (NIST), Tech. Rep. NIST.CSWP.02122014, 2014, Último acceso: 4 de marzo de 2025. [Online]. Available: http://nist.gov/cyberframework/

[3] C. I. Cybersecurity, "Framework for improving critical infrastructure cybersecurity," *URL: https://nvlpubs. nist. gov/nistpubs/CSWP/NIST. CSWP*, vol. 4162018, no. 7, 2018.

[4] V. M. Vilches, L. A. Kirschgens, A. B. Calvo, A. H. Cordero, R. I. Pisón, D. M. Vilches, A. M. Rosas, G. O. Mendia, L. U. S. Juan, I. Z. Ugarte, E. Gil-Uriarte, E. Tews, and A. Peter, "Introducing the robot security framework (rsf), a standardized methodology to perform security assessments in robotics," 2021. [Online]. Available: https://arxiv.org/abs/1806.04042

[5] C. M. Braga, M. A. Serrano, and E. Fernández Medina, "Hacia una propuesta de metodología para el desarrollo de proyectos de ia confiables," in *Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)*. Sevilla, España: Universidad de Sevilla. Escuela Técnica Superior de Ingeniería Informática, 2024, pp. 270–277.

[6] E. Basan, A. Gritsynin, and T. Avdeenko, "Framework for analyzing the security of robot control systems," in *2019 International Conference on Information Systems and Computer Science (INCISCOS)*, 2019, pp. 354–360.

[7] S. Oruma, R. Colomo-Palacios, and V. Gkioulos, "Architectural views for social robots in public spaces: business, system, and security strategies," *International Journal of Information Security*, vol. 24, no. 1, p. 12, November 2024. [Online]. Available: https://doi.org/10.1007/s10207-024-00924-x

[8] Comisión Europea, "Reglamento (UE) 2024/1689 sobre Inteligencia Artificial y Seguridad en Sistemas Autónomos," 2024, Último acceso: 19 de marzo de 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai

[9] Boletín Oficial del Estado, "Texto completo del Reglamento (UE) 2024/1689," 2024, Último acceso: 19 de marzo de 2025. [Online]. Available: https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

[10] M. Penmetcha, S. S. Kannan, and B.-C. Min, "Smart cloud: Scalable cloud robotic architecture for web-powered multi-robot applications," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 2397–2402.

[11] K. E. Chen, Y. Liang, N. Jha, J. Ichnowski, M. Danielczuk, J. Gonzalez, J. Kubiatowicz, and K. Goldberg, "Fogros: An adaptive framework for automating fog robotics deployment," in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2021, pp. 2035–2042.

[12] M. A. González-Santamarta, "yolo_ros," Feb. 2023. [Online]. Available: https://github.com/mgonzs13/yolo_ros

[13] ——, "whisper_ros," May 2023. [Online]. Available: https://github.com/mgonzs13/whisper_ros

[14] M. González and contributors, "llama_ros: Integration of llama with ros," GitHub, Último acceso: 19 de marzo de 2025. [Online]. Available: https://github.com/mgonzs13/llama_ros

[15] S. John, R. R. F. Del, K. Evgeniy, O. Helen, H. Idan, U. Kayla, H. Ken, S. Peter, A. Rakshith, B. Ron *et al.*, "Owasp top 10 for llm apps & gen ai agentic security initiative," Ph.D. dissertation, OWASP, 2025.

[16] OWASP, "OWASP LLM Top 10 - Principales riesgos de seguridad en modelos de lenguaje," 2024, Último acceso: 19 de marzo de 2025. [Online]. Available: https://genai.owasp.org/es/llm-top-10/

[17] Parlamento Europeo y Consejo de la Unión Europea, "Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Reglamento General de Protección de Datos, GDPR)," 2016, Último acceso: 19 de marzo de 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679

[18] N. I. of Standards and Technology, "Guide to intrusion detection and prevention systems (idps) (sp 800–94)," 2007, Último acceso: 4 de marzo de 2025. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-94.pdf

[19] F. Martín, E. Soriano-Salvador, J. M. Guerrero, G. Guardiola Múzquiz, J. C. Manzanares, and F. J. Rodríguez, "Towards a robotic intrusion prevention system: Combining security and safety in cognitive social robots," *Robotics and Autonomous Systems*, vol. 190, p. 104959, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889025000454

[20] P. A. Networks, "Next-generation firewalls," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.paloaltonetworks.com/network-security/next-generation-firewall.html

[21] ——, "Pan-os 10.2 administrator's guide: Firewall administration," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://docs.paloaltonetworks.com/pan-os/10-2/pan-os-admin/firewall-administration

[22] S. Inc., "Splunk enterprise security," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.splunk.com/en_us/software/enterprise-security.html

[23] R. S. LLC, "Rsa secure access and authentication suite," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.rsa.com/en-us/products/rsa-securid-access

[24] I. Okta, "Okta identity cloud," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.okta.com/learn/okta-identity-cloud/

[25] C. Systems, "Cisco trustsec," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.cisco.com/c/en/us/solutions/enterprise-networks/trustsec/index.html

[26] I. VMware, "Vmware nsx technical documentation," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://docs.vmware.com/en/VMware-NSX/index.html

[27] M. Corporation, "Configuration manager documentation," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://docs.microsoft.com/mem/configmgr/

[28] Ivanti, "Ivanti patch for endpoint manager," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.ivanti.com/products/patch-for-endpoint-manager

[29] Darktrace, "Ai cybersecurity platform," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.darktrace.com/products

[30] I. CrowdStrike, "Crowdstrike falcon platform," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.crowdstrike.com/platform/

[31] C. S. Ltd., "Cyberark identity security platform," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.cyberark.com/es/solutions/securing-cloud-operations/

[32] I. Trellix, "Acerca de trellix," 2025, Último acceso: 4 de marzo de 2025. [Online]. Available: https://www.trellix.com/es-es/about/

[33] G. Guardiola-Múzquiz and E. Soriano-Salvador, "Sealfsv2: combining storage-based and ratcheting for tamper-evident logging," *International Journal of Information Security*, vol. 22, no. 2, pp. 447–466, 2023.

XAI for safe AI: technical-legal misconceptions and misalignments

Marina Buitrago-Perez
Computer Security Lab (COSEC)
Universidad Carlos III de Madrid
Leganes, Spain
100428967@alumnos.uc3m.es

Jose Maria de Fuentes
Computer Security Lab (COSEC)
Universidad Carlos III de Madrid
Leganes, Spain
josemaria.defuentes@uc3m.es

Ofelia Tejerina
Observatorio LegalTech
Garrigues-ICADE
Madrid, Spain
otejerina@comillas.edu

Lorena Gonzalez-Manzano
Computer Security Lab (COSEC)
Universidad Carlos III de Madrid
Leganes, Spain
lgmanzan@inf.uc3m.es

Nicolas Anciaux
PETSCRAFT project-team
Inria, Université Paris-Saclay, UVSQ
Paris, France
nicolas.anciaux@inria.fr

Luis Ibanez-Lissen
Computer Security Lab (COSEC)
Universidad Carlos III de Madrid
Leganes, Spain
luibanez@pa.uc3m.es

***Abstract*—Building trusted AI models is of outmost relevance both from a societal and legal perspectives. In this vein, explainable AI (XAI) has been set as one of the technical cornerstones. For XAI to be successful, the *two perspectives*, namely regulators and technologists, must be aligned. However, this cannot be taken for granted considering the different evolution paces of both perspectives in the context of AI. This paper identifies a number of misalignments between legal and technical issues. Moreover, a survey, distributed to 31 practitioners of both perspectives, show a number of misconceptions, that is, technical and legal issues that are misleading or unknown by professionals, thus calling for a coordinated development for both perspectives in the short term.**

***Index Terms*—AI, explainable AI, safe AI, techniques, regulations, misalignments, misconceptions**

I. Introduction

Artificial Intelligence (AI) has become an integral component of modern technological ecosystems, driving innovations across diverse fields such as healthcare, finance, and autonomous systems. The increasing reliance on AI models, particularly those based on deep learning, has raised concerns about their robustness, fairness, and reliability. Many AI systems operate as black boxes, making it difficult to interpret their decision-making processes, which leads to significant risks in critical applications [1]. Security threats such as adversarial attacks [2] and data poisoning [3] have further exacerbated concerns regarding the safe deployment of AI. These vulnerabilities underscore the necessity for frameworks that enhance AI transparency and accountability.

Explainable AI (XAI) has emerged as a promising avenue to address the opaqueness of AI models by providing insights into their decision-making mechanisms [4]. The goal of XAI is to produce human-interpretable explanations of AI outputs, thereby increasing user trust and enabling better oversight of AI behavior. However, while XAI methods contribute to understanding AI behavior, they do not necessarily guarantee the trustworthiness of AI systems, as explanations may not be insighful enough.

The concept of safe AI goes beyond interpretability. It includes requirements such as robustness, fairness, accountability, and compliance with legal and ethical standards [5]. Ensuring AI safety requires a combination of technical safeguards, regulatory measures, and socio-technical considerations. Despite advancements in AI governance, there is a gap between explainability and actual AI safety, making it crucial to analyze the limitations of XAI in achieving safe AI.

Several prior works have explored the relationship between XAI and safe AI. [6] discuss different explainability techniques and their applicability in high-stakes environments, while [7] examine how XAI can aid in identifying adversarial vulnerabilities. Additionally, [8] investigate counterfactual explanations as a means of enhancing AI accountability, though their effectiveness in ensuring compliance with legal standards remains debated. Other studies, such as [9], argue that inherently interpretable models should be prioritized over post hoc explainability techniques to achieve true AI safety.

Motivation. Despite these efforts, there is still an unresolved technical-legal challenge in aligning XAI methodologies with the stringent requirements of safe AI, warranting further interdisciplinary research. However, this requires mutual knowledge between technologists and lawyers, which is yet to be proven. For example, the European AI Act requires human operators be able "to correctly interpret the high-risk AI system's output". However, from a technical viewpoint, interpreting, explaining and understanding are three different notions with diverse mechanisms and level of maturity. Therefore, successful development of safe AI models requires a common ground of understanding between all stakeholders.

This work has been supported by Grant PID2023-150310OB-I00 (MORE4AIO, funded by the Spanish Ministerio de Ciencia, Innovación y Universidades, the Spanish Agencia Estatal de Investigación and FEDER funds

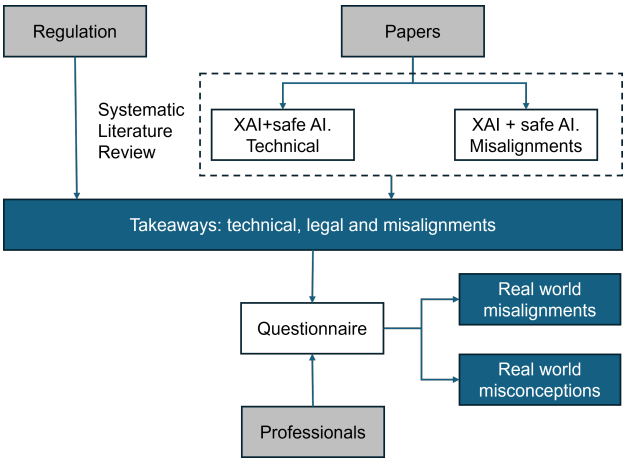


Fig. 1: Methodology

Contribution. In this paper, we analyse the degree of mutual knowledge between these communities. This helps us identify a number of *misalignments*, which unveil the fields in which more joint cooperation is required. In the same vein, we identify a number of *misconceptions*, that is, notions that are misleading or unknown. These findings pave the way for the creation of a common technical background to support future safe AI developments.

More precisely, our contributions are as follows:

C1. We identify a number of takeaways that summarize (1) the state of the art of XAI for safe AI from a technical perspective, (2) the underlying regulations and (3) the misalignments between these two areas. Although some papers have addressed these issues [10]–[13], they did not provide with concrete items to work on.

C2. We measure the degree of (lack of) knowledge in the technical and legal communities regarding the identified takeaways. In this regard, we conduct a survey among 31 professionals. Although they mostly belong to the European Union (Spain and France), it also includes citizens from China and Argentina. This leads us to identify misalignments and misconceptions that deserve further training/upskilling, technical or legal characterization.

The remaining of the paper is structured as follows. Section II describes the methodology. Sections III, IV and V show the identified takeaways concerning technical and legal aspects of XAI for safe AI, as well as the misalignments between these two, respectively. Section VI measures the actual misconceptions and misalignments considering real-world professionals. Section VII describes the related work. Lastly, Section VIII concludes the paper and points out future work directions.

II. Methodology

The methodology followed in this paper is composed by two main steps, as depicted in Figure 1.

Identification of technical, legal and misalignments between the two (Figure 1, upper part).

We conducted a literature review of major academic databases through Google Scholar from 2019 to the present. Some of the most important search terms were: ”XAI”, “XAI Survey”, “Explainable AI”, ”XAI limitations”, ”AI transparency”, ”AI interpretability”, ”Safe AI”, ”XAI security threats”, ”XAI attacks”, “safe AI”, “Secure explainability”. After manual inspection of 117 papers, 39 papers were finally considered as relevant. This review synthesized the current state of XAI research within both legal regulations and technical spheres. These findings exposed a notable gap between technological and regulatory concerns.

To further explore this gap, we searched for articles published from 2020 onward that explicitly address the intersection of these two domains. Using the query terms ”XAI” and “GDPR”, ”XAI” and “AI Act”, ”XAI” and “legal”, ”Data protection in XAI”, ”XAI” and “trade secrets”, ”XAI legal challenges”, “XAI” and (“misalignment” OR “misconception”), “privacy” and “XAI”, “regulation” and “XAI”, “XAI” and “knowledge gaps”, 52 papers were manually reviewed to finally take 13 papers into consideration.

Measuring real-world misconceptions and misalignments (Figure 1, lower part). The takeaways collected in the previous phase are useful in describing the status of XAI for safe AI from a theoretical perspective. However, it lacks real-world vision. For this purpose, we developed a targeted survey aimed at gathering insights from technologists and legal experts working with AI. The necessity of this survey became clear as our literature analysis highlighted a critical lack of multidisciplinary understanding and the poor exchange of information between these two key groups. Thus, it helped us identifying important real-world misconceptions and misalignments to further inspire a research roadmap.

III. XAI for Safe AI. Technical State of the Art

This section provides with an overview of existing techniques, to later identify a number of takeaways.

A. Overview of techniques

XAI aims to provide a layer of transparency to deep learning models. Since these models, including LLMs are extremely complex, providing human-understandable explanations is relevant to build a certain degree of trust in their operation [14].

Various XAI techniques, including feature attribution methods, surrogate models, and counterfactual explanations, have been proposed to improve interpretability [15]. On the one hand, model-agnostic methods, such as SHAP [14] and LIME [15], generate explanations applicable across different AI models. On the other hand, model-specific techniques leverage internal structures, such as attention mechanisms in transformers, to provide insights. As an example, linear probing is a commonly used technique for evaluating the knowledge encoded in the internal representations of a model [16]. It involves training a relatively simple linear classifier, such as a logistic regression model, using features extracted from the model. Among these features, internal activations are a

promising approach. They are the intermediate representations of data as it passes through the model layers [17].

Despite the interest of these techniques, they currently face two main limitations. On one side, currently interpretability comes at the cost of accuracy. Thus, when models expend a lot of internal resources in providing explanations, they sacrifice their predictive power [18]. On the other side, the relative immaturity of these technologies make them be potentially inconsistent. Therefore, explanations from different XAI techniques may not be compatible themselves.

B. Takeaways

Tech1 **SHAP and LIME are the most popular XAI techniques**. SHAP uses game theory to compute the impact of each input feature whereas LIME works by building simple models around individual predictions to show which features are more relevant in each case. Grad-CAM [19]–[24] is frequently used in healthcare applications, offering visual explanations of neural network decisions. DeepLIFT [19], [20], [25], [26] provides insights into deep learning models by tracking neuron activations. Counterfactual explanations [11], [12], [18], [19], [25]–[28] are widely applied in legal and financial AI systems to explain decision-making by presenting alternative scenarios. Other popular methods include DALEX [10], [18], [23] and Anchors [18], [23], [29], which enhance interpretability in machine learning models.

Tech2 **Lack of XAI benchmarks**. Explainability remains ill-defined due to the absence of universally accepted metrics, evaluation protocols, or agreed-upon benchmarks [10]–[12], [19]–[23], [25]–[28], [30], [31]. Researchers struggle to measure, compare, or even define "interpretability" consistently, hindering reproducibility and objective assessment of different XAI approaches. Multiple metrics are proposed, but their utility is not standardized. Without a concrete evaluation framework explainability remains unevenly implemented and difficult to benchmark across domains and teams.

Tech3 **XAI Increases the attack surface of AI systems** XAI introduces new security vulnerabilities. Explainability methods can be leveraged by adversaries to reverse-engineer AI models, leading to model extraction attacks. Additionally, explanations can reveal biases that attackers may exploit to manipulate model behavior. Studies emphasize the need for robust defense mechanisms to mitigate these risks, but the field lacks standardized security solutions for XAI [11], [20], [24], [26], [28], [31]–[33].

Tech4 **Limits of Interpretable models and challenges in scaling XAI** High-performance AI architectures, especially deep learning, often lack intuitive transparency, and simplifying them for explainability may reduce accuracy. Local explanation methods are not always trustworthy for complex, evolving systems, and scaling these approaches to large datasets or real-time environments is computationally intensive [11], [18]–[23], [25], [26], [28], [32], [34], [35].

Tech5 **Biases impose a challenge to XAI.** Researchers stress that explanations must reveal potential discriminatory factors and enable robust fairness assessments, but standardized bias detection and mitigation techniques remain insufficient [11], [22], [24], [26], [30]–[32], [35].

Tech6 **XAI understandability**. Effective XAI hinges on producing understandable explanations for non-technical stakeholders, such as clinicians or financial analysts, who risk over-reliance on or mistrust of AI if explanations are opaque or overly complex [10], [11], [20], [25], [32], [34], [35]. Automation bias can emerge when superficial justifications lull users into uncritical acceptance of AI outputs. Conversely, overly detailed technical explanations can deter adoption. Achieving clarity, brevity, and domain relevance is crucial to fostering trust, proper reliance, and informed oversight of AI systems.

Tech7 **Need for domain-specific XAI.** Close collaboration between domain experts and AI developers is essential to ensure that explanations are meaningful, legally compliant, and technically feasible [11], [12], [29], [35]. Because one-size-fits-all solutions are ineffective, sustained interdisciplinary efforts are necessary to accommodate each field's data characteristics, legal context, and specific interpretability objectives.

Tech8 **A few domains are attracting most attention concerning XAI.** Healthcare remains the most extensively studied domain, with significant emphasis on improving transparency in medical decision-making and diagnostics [11], [21]–[25], [30], [36], [37]. Other well-explored domains include finance [11], [18], [24], [25], [30], [36], law [12], [25], [27], [30], [35], cybersecurity [11], [26], [31], [33], and IoT [19], [23], along with Edge AI [19] and autonomous systems [11], [19], though to a lesser extent.

IV. Legal issues related to XAI and safe AI

In this section, the impact of XAI in two main AI-related issues are described – the automated decision making leveraging personal data, and personal data protection and trade secrets. Afterwards, a number of takeaways are identified.

It must be noted that while regulations do not typically opt for the concept of "safe AI", they impose a number of requirements to meet a minimum level of trustworthiness.

A. XAI and the right not to be subject to a decision based solely on automated processing under GDPR

In the European Regulatory framework, as set forth in art. 22(1) General Data Protection Regulation (hereinafter, GDPR) "[t]he data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her".

The recent judgment by the Court of Justice of the European Union (CJEU) in case C-634/21 [38] sheds light on how

automated processes, such as credit scoring, fall within the scope of Article 22(1) if they significantly affect individuals and if the outcomes heavily influence subsequent decisions by third parties. The ruling underscores the requirement for meaningful human oversight in automated profiling processes: even when a human ostensibly grants final approval, any automated component that decisively shapes the outcome may be captured by the prohibition in art. 22(1) if the purported human review is merely symbolic rather than substantive.

Given the importance of ensuring human oversight is not merely symbolic but genuinely influential, regulatory frameworks stress the need for transparency in its implementation. To this end, the Spanish Data Protection Agency mandates that, as part of the Data Protection Impact Assessment, data controllers must identify and document both the degree of human intervention and the specific point at which such intervention occurs during the decision-making process [39].

Extending these principles, the European Union Artificial Intelligence Act (hereinafter, AI Act) also underscores the importance of human oversight in art. 14(4)(c) AI Act, particularly for high-risk AI systems. It mandates that, where appropriate and proportionate, human operators must be able "to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available". This requirement has propelled XAI into the spotlight, as these techniques offer an additional layer of transparency to "black-box" models [40]. By clarifying the reasoning behind automated decisions—such as how input features are weighted or which logical pathways lead to a given conclusion—XAI establishes a clearer, more comprehensible link between inputs and outputs [41]. In turn, it facilitates the substantive human review that both the GDPR and the emerging AI regulatory framework increasingly demand.

B. Implications of XAI for data protection, privacy and trade secrets

Art. 5(1)(f) GDPR imposes the duty to process personal data "[. . .] in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures." From this provision, it follows that any AI system handling personal data must be equipped with robust safeguards to deter improper access or exposure of sensitive information.

In turn, art. 5(2) GDPR introduces the principle of accountability. It implies adopting security policies as well as demonstrating their effectiveness. These mechanisms (arts 25 and 32 GDPR) must be in line with the state of the art considering the risks at stake.

Consistent with this regulatory approach, in art. 15(5) AI Act it is stipulated that "[h]igh-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities" and that "[t]he technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws."

As analyzed above, XAI systems exhibit notable vulnerabilities when compared to traditional black-box models, since the interpretability that benefits users may likewise facilitate adversarial attacks. For instance, methods such as SHAP, by disclosing critical information about a model's internal logic, can inadvertently enable attackers to identify key features and decision thresholds, refining their strategies for evading detection in security systems. Similarly, the use of LIME to interpret black-box models expands the attack surface by generating local explanations via surrogate models, thus revealing the original model's decision boundaries, enabling targeted attacks and gradient estimation. Furthermore, LIME can be exploited to extract proprietary models, compromising intellectual property, or to poison model explanations or manipulate feature importance, thereby distorting security assessments and concealing system vulnerabilities.

Therefore, although XAI aligns with the transparency principle by allowing data subjects to seek explanations of automated decisions, it may also jeopardize the confidentiality requirement of arts. 5(1)(f), 25 and 32 GDPR, as well as the protection requirements imposed by art 15(5) AI Act.

C. Takeaways

Law1 **Regulatory conflicts between transparency, data protection and intellectual property and trade secrets**. While European regulations impose strict data privacy safeguards, they simultaneously demand increasing model explainability. This dual requirement generates friction, particularly in sectors where disclosing AI decision rationales risks inadvertently exposing sensitive data [10]–[12], [22], [26], [27], [29], [32], [35], [42]. The lack of established legal doctrine or jurisprudence further exacerbates this challenge, leaving organizations without clear precedents to navigate these conflicting obligations. Moreover, the transparency mandate threatens enterprises' ability to safeguard trade secrets and intellectual property—an issue that remains largely overlooked in academic discourse [12], [27], [32].

Law2 **Lack of international legal frameworks for AI and XAI**. There is no comprehensive global framework governing AI and the expected standards in explainability. The fragmented nature of AI regulations worldwide results in inconsistencies in implementation, making compliance challenging for organizations operating across multiple jurisdictions. Researchers have pointed out that without international consensus on what constitutes explainability and how it should be enforced, companies and developers are left to navigate a complex and often contradictory regulatory landscape [25], [27], [35]

Law3 **Secure explainability remains elusive under legal constraints** Balancing transparency with security is complex,

as elucidating AI decisions may introduce vulnerabilities like model inversion attacks, which can compromise user privacy or leak proprietary information. Current XAI methods often fail to satisfy both legal transparency demands and protective measures simultaneously. This predicament is especially pronounced in sensitive domains where compliance with data protection laws and ensuring robust model security converge, making truly "secure explainability" an unsettled goal [12].

V. Identifying technical-legal misalignments

The literature underscores a profound gap between legal and technical understandings of AI transparency, resulting in regulatory ambiguity, unrealistic expectations [43], [44], and challenges in implementation [35]. Legal frameworks such as the AI Act and GDPR mandate transparency yet fail to delineate its precise scope, leaving AI developers without clear compliance guidelines to operationalize legal mandates, whereas policymakers often underestimate the technical constraints inherent in model architectures [10], [12], [45]. The misalignment between XAI methodologies and legal requirements further complicates certification and oversight, as regulators tend to assume transparency can be achieved through simplified explanations for end-users, whereas researchers emphasize the necessity of advanced interpretability methods [46]. This disciplinary divide stresses the urgency of interdisciplinary collaboration to establish meaningful and practicable transparency standards [13], [27], [47]–[49].

A. Takeaways

Misal1 **Regulatory vagueness and incomplete guidance on AI transparency** Multiple studies stress that both the GDPR and the AI Act impose transparency obligations yet fail to specify what level of openness is considered "sufficient" or "appropriate" [12], [35], [43], [44]. This lack of explicit criteria intensifies the challenges faced by AI developers, who must reconcile legal mandates with the absence of standardized metrics for algorithmic explainability, all while contending with safety requirements.

Misal2 **Persistent legal–technical mismatch in defining "transparency"** In both legal and technical discourse, "transparency" is conceptualized markedly different—legal experts prioritize accountability and user comprehension, whereas developers concentrate on elucidating how models process inputs [10], [13], [35], [44], [48]. This discrepancy results in divergent objectives: regulators expect thorough yet accessible explanations, while AI practitioners must navigate sophisticated model architectures that defy any simplistic, "one-size-fits-all" approach to explainability.

Misal3 **Unrealistic legislative expectations about XAI capabilities** Several authors argue that policymakers overestimate XAI's capacity to produce intuitive, user-friendly explanations for every context [10], [44], [45], [47]. By treating XAI as an all-encompassing solution—rather than a range of context-dependent methods and approaches—lawmakers risk imposing obligations that outstrip current technological limits, thereby complicating implementation efforts and straining developers' capacity to meet regulatory standards.

VI. Measuring misalignments and misconceptions in practice

All previous findings come from the scientific domain. Therefore, it is necessary to confirm the current status of real-world professionals. This section focuses on this matter. First, the structure of a questionnaire distributed to professionals is presented. Afterwards, the profile of respondents is described. Lastly, the identified common grounds between both communities, their misconceptions and misalignments are presented. Figure 2 illustrates some relevant results of the survey.

A. Survey structure

The online survey can be found in Google Forms[1]. The survey includes a balanced set of technical and legal questions. Technical questions were aimed to assess the baseline knowledge of jurists regarding core AI concepts, ensuring their ability to legislate and guide regulatory practices effectively. Conversely, legal questions were conceived to determine whether technologists have an adequate grasp of the regulatory requirements that are necessary for developing AI systems in compliance with the legal framework. All questions were aligned with the discovered takeaways, as shown in Table I.

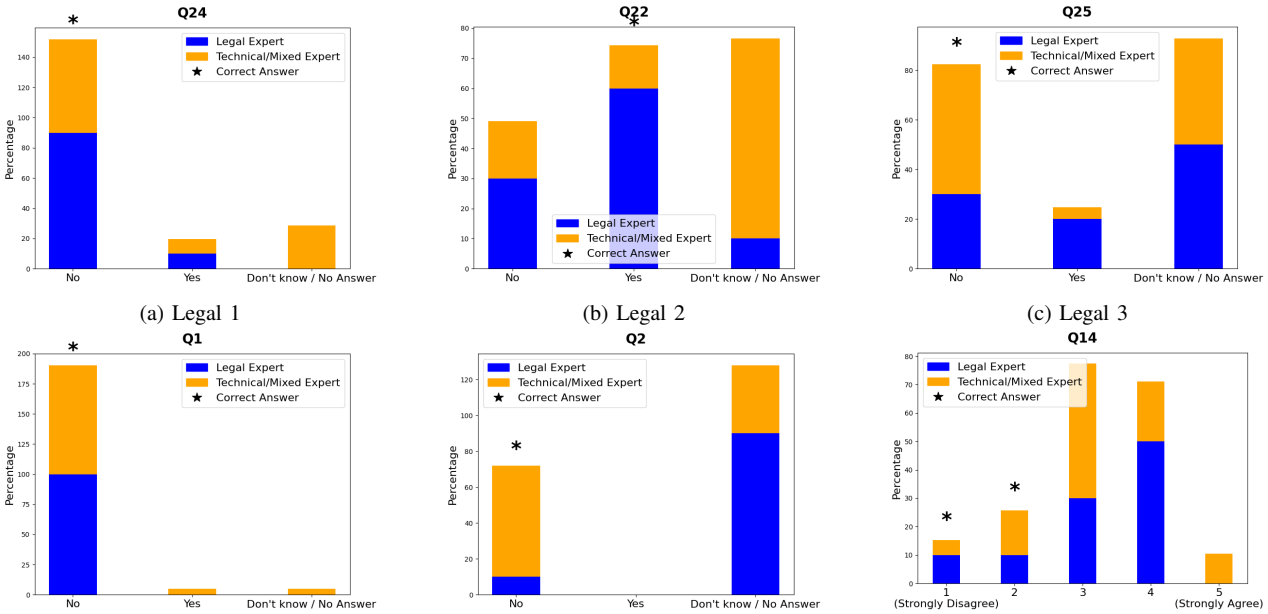
TABLE I: Coverage of each takeaway in the survey

Questions	Tech1	Tech2	Tech3	Tech4	Tech5	Tech6	Tech7	Tech8	Law1	Law2	Law3	Misal1	Misal2	Misal3
Q1		✓								✓				
Q2	✓			✓										✓
Q3				✓										
Q4			✓											
Q5					✓									
Q6						✓			✓					✓
Q7								✓						
Q8				✓										
Q9							✓							
Q10						✓	✓							
Q11		✓												
Q12				✓										✓
Q13						✓								
Q14					✓	✓								
Q15			✓											
Q16									✓					
Q17									✓					
Q18										✓				
Q19										✓				
Q20	✓													
Q21														✓
Q22												✓		
Q23											✓	✓	✓	
Q24		✓							✓	✓		✓		
Q25									✓		✓			✓
Q26		✓							✓	✓		✓		
Q27									✓		✓	✓		
Q28		✓									✓	✓		✓
Q29									✓					
Q30									✓					
Q31		✓											✓	
Q32									✓		✓	✓		

The survey included a mix of question types, including true/false and Don't Know/No Answer and Likert scale questions. The use of Likert scales allowed us to assess the intensity of respondents' opinions, while true/false and "don't know" responses provided a clearer indication of misunderstandings.

All survey questions have a correct answer derived from the analyzed state of the art or directly from regulations. In yes/no questions, selecting "no" when the correct answer is "yes" (or vice versa) is considered a misconception. If a significant number of respondents choose "don't know" in a particular area, we interpret this as a misalignment—if there is

[1]https://forms.gle/VVKDdBsEMHwb8aZe8 , last access March 2025.



(a) Legal 1 (b) Legal 2 (c) Legal 3

(d) Tech1 (e) Tech 2 (f) Tech 3

Fig. 2: Excerpt of survey results

discordance between technologists and jurists. The same logic applies to Likert scale questions – when the correct response leans toward "strongly disagree," we consider 1 and 2 correct, whereas when it leans toward "strongly agree", 4 and 5 are correct. Responses of 3 are categorized as "don't know".

B. Respondents profile

Participants were selected among the network of contacts of the authors, including academic researchers and lawyers. 67% of respondents are identified as technical experts, while about 33% as legal experts. All participants exhibited a high level of specialization in AI development, cybersecurity, AI governance, or technology law, holding roles such as Data Protection Officers, professionals in data protection agencies, LLM/NLP developers, AI cybersecurity professionals or AI/tech law practitioners. In terms of academic credentials, 36% hold a PhD, 45% a master's degree, and 16% a bachelor's degree as their highest qualification.

C. Common ground between technologists and legal professionals

SC1 **Lack of universal XAI standards** Both groups agree that no universally accepted standards exist for measuring AI explainability (Q1). This lack of a unified benchmark makes it difficult to compare different XAI approaches.

SC2 **Uncertainty about scalability and real-time use** Participants also share uncertainties regarding whether these explainability methods can effectively scale to large datasets or be applied in real-time scenarios (Q12), highlighting a broader knowledge gap in practical deployment.

SC3 **Incomplete and uncertain legal frameworks** Respondents consistently concur that neither a global nor an EU-level framework definitively outlines AI transparency obligations (Q24). This regulatory patchwork leaves organizations navigating evolving or incomplete guidelines.

D. Identified misconceptions

SMc1 **"Increasing explainability inevitably reduces predictive accuracy"** In Q3, it can be observed that opinions were sharply divided over whether enhancing transparency necessarily compromises model performance. This split suggests lingering uncertainty about the real trade-offs between explainability and accuracy.

SMc2 **"A clear, interpretable explanation implies the AI model is reliable"** As seen in Q14, many legal professionals overestimate models' trustworthiness based solely on its explainability. Technical experts, however, are more cautious, recognizing that transparency alone does not guarantee the model's correctness or robustness.

SMc3 **"Transparency requirements do not conflict with data protection, and current XAI fully ensures security"** This perspective, predominantly held by legal experts, encompasses two misconceptions. First, they perceive no genuine conflict between transparency mandates and data protection laws (Q16), overlooking potential risks in exposing sensitive data or model details. Second, according to Q25 answers, some respondents presume that existing XAI methods already satisfy both security and transparency obligations, thus highlighting the need for greater awareness within the legal sector.

E. Identified misalignments

SMa1 **Uncertainty over the adequacy of common XAI methods in Deep Learning** A significant number of respondents—both legal and technical—remain unsure whether widely used explainability methods (e.g., SHAP, LIME) truly suffice for deep learning models. For instance, Q2 indicates that 90% of legal experts and 40% of technical experts expressed uncertainty. This reflects a broader lack of clarity about how and to what extent these methods effectively enhance AI transparency.

SMa2 **Regulatory ambiguities in GDPR and the AI Act** There is notable confusion regarding regulations such as the GDPR's "right to explanation" (Q22) and the EU's AI Act (Q28). Legal experts do not uniformly agree on how these requirements should be implemented, and technical experts often show limited familiarity with key legal texts or case law (Q27). This gap complicates efforts to reconcile transparency and security obligations, as well as questions around liability for non-compliance. Uncertainty about the AI Act's specifics, in particular, underscores a deeper misalignment between legal and technical perspectives.

SMa3 **Cross-disciplinary disconnect on defining transparency** Technologists and legal professionals define "transparency" in different ways, as indicated in Q31. Many participants (Q21) also note that legal professionals often lack in-depth technical understanding. Consequently, these differences in viewpoint make it harder to develop shared strategies for AI transparency.

VII. Related work

An overview of related literature is given in Table II, showing the coverage of technical and legal challenges, as well as any noted misalignments and misconceptions.

The identification of technical issues surrounding XAI for safe AI has been extensively addressed. Most of these studies focus on widely used explanation methods (e.g. [26], [30]) and their application in a few domains that are attracting most interest (v. gr. [21], [22], [27], [35]), yet devote comparatively less attention to secure explainability and the introduction of new security vulnerabilities (e.g. [24], [31], [33]).

Similarly, certain legal aspects have received substantial research attention. For instance, in [42] the tension between mandated transparency and the opacity of large-scale models is highlighted. In addition, much of the literature featuring legal aspects also addresses misalignments (e.g. [47]–[49]) such as ambiguous regulatory language [35] and divergent definitions of "transparency" across disciplines [44], while offering limited insight into practical strategies that reconcile security requirements with transparency obligations [12]. This underscores a notable deficit in actionable guidance.

Interestingly, none of these previous studies surveyed both technologists and jurists to evaluate real-world knowledge gaps, misconceptions, and cross-domain frictions between technical and legal fields. Consequently, they do not capture the actual level of expertise among practitioners. The present work fills this gap, thereby providing a more comprehensive perspective on the current state of XAI in practice.

TABLE II: Related work

Ref	Year	Technical issues	Legal issues	Misalignments	Misconceptions
[28]	2020	✓	-	-	
[45]	2021	-	✓	✓	
[27]	2022	-	✓	✓	
[26]	2022	✓	✓	-	
[20]	2022	✓	-	-	
[43]	2022	-	✓	✓	
[30]	2023	✓	-	-	
[11]	2023	✓	✓	✓	
[18]	2023	✓	-	-	
[33]	2023	✓	-	-	
[44]	2023	-	✓	✓	
[49]	2023	-	✓	✓	
[10]	2024	✓	✓	✓	
[25]	2024	✓	✓	-	
[19]	2024	✓	-	-	
[34]	2024	✓	-	-	
[32]	2024	✓	✓	-	
[21]	2024	✓	-	-	
[22]	2024	✓	✓	-	
[24]	2024	✓	-	-	
[12]	2024	✓	✓	✓	
[42]	2024	-	✓	-	
[29]	2024	✓	✓	-	
[31]	2024	✓	-	-	
[13]	2024	✓	✓	✓	
[46]	2024	✓	-	✓	
[47]	2024	-	✓	✓	
[48]	2024	-	✓	✓	
[36]	2025	✓	-	-	
[37]	2025	✓	-	-	
[23]	2025	✓	-	-	
[35]	2025	-	✓	✓	
OURS	2025	✓	✓	✓	✓

VIII. Conclusion

The growing relevance of AI models in diverse scenarios, including high-risk ones, calls for the adoption of safety measures in their development and operation. However, the notion of safe AI may be dramatically different from the technical and legal perspectives. In the same vein, explainability (XAI) can take different roles and relevance levels to achieve safe AI, depending on the extent, purpose and expectations put into it. This paper has analysed the current status of the technical and legal aspects of XAI for safe AI. Beyond identifying several takeaways, a number of misconceptions and misalignments have been discovered.

As future work, our findings are helpful for a roadmap of further research that should tackle training/upskilling as well as technical and legal coordinated advances. The creation of a methodological framework for such a coordination, together with the standardization of XAI, are also future directions.

References

[1] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2014.

[3] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[4] A. B. A. et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[5] D. A. et al., "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[7] K.-R. M. Wojciech Samek, Thomas Wiegand, "Towards explainable artificial intelligence," *arXiv preprint arXiv:1904.12479*, 2019.

[8] C. R. Sandra Wachter, Brent Mittelstadt, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law Technology*, vol. 31, pp. 841–887, 2017.

[9] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[10] P. Biecek and W. Samek, "Explain to question not to justify," *arXiv preprint*, 2024.

[11] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.

[12] B. Fresz, E. Dubovitskaya, D. Brajovic, M. F. Huber, and C. Horz, "How should ai decisions be explained? requirements for explanations from the perspective of european law," in *AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, October 2024, pp. 438–450.

[13] K. A., S. D., A. E., and B. S., "A legal framework for explainable artificial intelligence," *Center for Law Economics Working Paper Series*, vol. 9, 2024.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[15] S. S. Marco Tulio Ribeiro and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[16] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *International Conference on Learning Representations (ICLR) Workshops*, 2016.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[18] N. Rane, S. Choudhary, and J. Rane, "Explainable artificial intelligence (xai) approaches for transparency and accountability in financial decision-making," *SSRN*, 2023, available at SSRN 4640316.

[19] O. Vermesan, V. Piuri, F. Scotti, A. Genovese, R. D. Labati, and P. Coscia, "Explainability and interpretability concepts for edge ai systems," in *Advancing Edge Artificial Intelligence*. River Publishers, 2023, pp. 197–227.

[20] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of xai-based model improvement," *Information Fusion*, vol. 92, pp. 154–176, April 2023.

[21] C. Divakar, R. Harsha, K. Radha, D. V. Rao, N. Madhavi, and T. Bharadwaj, "Explainable ai for cnn-lstm network in pcg-based valvular heart disease diagnosis," in *2024 14th International Conference on Cloud Computing, Data Science Engineering (Confluence)*. IEEE, January 2024, pp. 92–97.

[22] A. A. Biswas, "A comprehensive review of explainable ai for disease diagnosis," *Array*, vol. 100345, 2024.

[23] R. Mariappan, "Extensive review of literature on explainable ai (xai) in healthcare applications," *Recent Advances in Computer Science and Communications*, vol. 18, no. 1, p. E200324228159, 2025.

[24] U. S. Kathait, A. Rana, R. Chauhan, and R. Rawat, "A comprehensive review of interpretability in ai and its implications for trust in critical applications," in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*. IEEE, October 2024, pp. 1683–1693.

[25] H. DeSimone and M. Leon-Espinosa, "Explainable ai: The quest for transparency in business and beyond," in *2024 7th International Conference on Information and Computer Technologies (ICICT)*. IEEE, March 2024, pp. 532–538.

[26] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. St. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 852–866, December 2022.

[27] B. Kuzniacki, M. Almada, K. Tylinski, Gorski, B. Winogradska, and R. Zeldenrust, "Towards explainable artificial intelligence (xai) in tax law: the need for a minimum legal standard," *World tax journal*, vol. 14, pp. 573–616, 2022.

[28] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[29] Górski, B. Kuźniacki, M. Almada, K. Tyliński, M. Calvo, P. M. Asnaghi, and J. I. Nigrelli, "Exploring explainable ai in the tax domain," *Artificial Intelligence and Law*, pp. 1–29, 2024.

[30] G. P. Reddy and Y. P. Kumar, "Explainable ai (xai): Explained," in *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. IEEE, April 2023, pp. 1–6.

[31] N. Rodríguez-Barroso, M. Luzón, and F. Herrera, "Rab ^{2}-def: Dynamic and explainable defense against adversarial attacks in federated learning to fair poor clients," *arXiv preprint*, 2024.

[32] T. Zeiser, D. Ehret, T. Lutz, and J. Saar, "Explainable ai in manufacturing," in *2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)*. IEEE, June 2024, pp. 1–8.

[33] J. Jagannathan, N. Labhade-Kumar, R. Rastogi, M. V. Unni, and K. K. Baseer, "Developing interpretable models and techniques for explainable ai in decision-making," *The Scientific Temper*, vol. 14, no. 04, pp. 1324–1331, 2023.

[34] J. Govea, R. Gutierrez, and W. Villegas-Ch, "Transparency and precision in the age of ai: evaluation of explainability-enhanced recommendation systems," *Frontiers in Artificial Intelligence*, vol. 7, p. 1410790, 2024.

[35] M. Almada, "Technical ai transparency: A legal view of the black box," 2025.

[36] M. Chakraborty, "Explainable artificial intelligence (xai): A perspective," in *International Ethical Hacking Conference*. Springer Nature Singapore, March 2024, pp. 47–63.

[37] O. R. Cawiding, S. Lee, H. Jo, S. Kim, S. Suh, E. Y. Joo, and J. K. Kim, "Symscore: Machine learning accuracy meets transparency in a symbolic regression-based clinical score generator," *Computers in Biology and Medicine*, vol. 185, p. 109589, 2025.

[38] "Judgment of the court of justice of the european union (first chamber) of december 7, 2023, case c-634/21, oq v. land hessen and schufa holding ag," 2023.

[39] S. D. P. Agency. Evaluation of human intervention in automated decisions. Spanish Data Protection Agency. [Online]. Available: https://www.aepd.es/prensa-y-comunicacion/blog/evaluacion-de-la-intervencion-humana-en-las-decisiones-automatizadas

[40] F. Ezzedine, "Privacy implications of explainable ai in data-driven systems," 2024, available at: https://arxiv.org/abs/2406.15789.

[41] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, p. 47, 2024.

[42] T. Cantens, "How will the state think with chatgpt? the challenges of generative artificial intelligence for public administrations," *AI SOCIETY*, pp. 1–12, 2024.

[43] M. Busuioc, D. Curtin, and M. Almada, "Reclaiming transparency: contesting the logics of secrecy within the ai act," *European Law Open*, vol. 2, pp. 1–27, 2022.

[44] B. Gyevnar, N. Ferguson, and B. Schafer, "Bridging the transparency gap: What can explainable ai learn from the ai act?" in *ECAI 2023*. IOS Press, 2023, pp. 964–971.

[45] A. Vorras and L. Mitrou, "Unboxing the black box of artificial intelligence: Algorithmic transparency and/or a right to functional explainability," in *EU Internet Law in the Digital Single Market*. Springer International Publishing, 2021, p. 247–264.

[46] B. Fresz, V. P. Göbels, S. Omri, D. Brajovic, A. Aichele, J. Kutz, and M. F. Huber, "The contribution of xai for the safe development and certification of ai: An expert-based analysis," *arXiv preprint*, 2024.

[47] Y. Walter, "Managing the race to the moon: Global policy and governance in artificial intelligence regulation—a contemporary overview and an analysis of socioeconomic consequences," *Discover Artificial Intelligence*, vol. 2, p. 14, 2024.

[48] N. Moreira, P. M. Freitas, and P. Novais, "Towards transparent ai: How will the ai act shape the future?" in *EPIA Conference on Artificial Intelligence*. Springer Nature Switzerland, 2024, pp. 296–307.

[49] B. Schütte, "Ai regulation in the eu: The future interplay between frameworks," in *YSEC Yearbook of Socio-Economic Constitutions 2023: Law and the Governance of Artificial Intelligence*. Springer Nature Switzerland, 2023, pp. 17–49.

Definiendo el concepto de explicación: propuesta de una taxonomía para explicabilidad en robótica

David Sobrín-Hidalgo
Grupo de Robótica
Universidad de León
Campus de Vegazana, 24071 León
dsobh@unileon.es

Ángel Manuel Guerrero-Higueras
Grupo de Robótica
Universidad de León
Campus de Vegazana, 24071 León
am.guerrero@unileon.es

Francisco Javier Rodríguez-Lera
Grupo de Robótica
Universidad de León
Campus de Vegazana, 24071 León
fjrodl@unileon.es

Vicente Matellán-Olivera
Grupo de Robótica
Universidad de León
Campus de Vegazana, 24071 León
vmato@unileon.es

***Resumen*—La explicabilidad de robots autónomos o eXplainable Autonomous Robots (XAR) en inglés es un campo de estudio que ha captado la atención de la comunidad investigadora en los últimos años, debido a su importancia no solo en la robótica sino también la seguridad, necesaria al tratarse interacciones constantes y directas entre humanos y máquinas. Sin embargo, el concepto de explicabilidad es complejo debido a su abstracción. Por esta razón, no existe un acuerdo establecido a la hora de definir el concepto de explicación. En el presente artículo se pretende afrontar este problema mediante la creación metodológica de una taxonomía. Proceso que no busca generar solo una clasificación jerárquica de conceptos relacionados, sino también alcanzar las definiciones de los mismos. A lo largo de este trabajo se presentan las definiciones propuestas y la clasificación planteada, con el objetivo de servir de punto de partida las nuevas investigaciones en el campo del XAR.**

***Index Terms*—Explicabilidad, taxonomía, terminología, robótica, XAR**

Tipo de contribución: *Investigación original*

I. Introducción

La creciente integración de la robótica en entornos HRI (*Human-Robot Interaction*) [1] ha impulsado la necesidad de aumentar la confianza entre máquina y usuario [2], y para ello es necesario que el comportamiento de los robots no solo sea transparente, sino también comprensible. Esta necesidad adquiere, por supuesto, un carácter imperativo cuando dichos entornos son críticos cómo la medicina, la conducción autónoma o los entornos industriales.

La explicabilidad en robótica es conocida por el acrónimo XAR (eXplainable Autonomous Robot) [3] y busca dotar a los robots autónomos de la capacidad de autoexplicar sus procesos de toma de decisiones. Dicha capacidad resulta indispensable a la hora de fomentar la confianza del humano hacia el robot, y así mejorar la percepción que los usuarios tienen sobre los procesos internos del sistema que en muchas ocasiones pueden actuar cómo ”cajas negras” para ellos.

Cómo es de esperar, la explicabilidad y la ciberseguridad están estrechamente ligadas ([4], [5]), ya que la comprensibilidad de un sistema facilita la detección de problemas e incidencias. Además la trazabilidad inherente a un sistema de explicabilidad robusto puede ser de utilidad para tareas relacionadas con el análisis forense, tal y cómo se ha demostrado en un trabajo previo [6].

No obstante, a pesar de la importancia de la explicabilidad, el estudio de XAR cuenta con un importante obstáculo conceptual: la ausencia de una terminología unificada. En concreto, el término explicación es complejo y abstracto. Es por esa razón que a través de la literatura se pueden encontrar diferentes definiciones según el autor.

Para afrontar dicho problema, este artículo propone una taxonomía para explicabilidad, es decir, una clasificación jerárquica que permitirá establecer las partes que componen una explicación y definir los conceptos principales. A la hora de afrontar esta investigación se ha seguido una metodología para asegurar la consistencia en cada una de sus fases, y cómo paso previo se ha establecido el propósito de la misma, el cuál permite acotar su alcance y dimensiones. El propósito definido se divide en los siguientes aspectos:

- **Usuario esperado:** Investigadores y organizaciones que trabajan en la generación de explicaciones en robots.
- **Uso esperado:** diseño y generación de explicaciones sobre el comportamiento de un robot.
- **Propósito:** proporcionar una clasificación de explicaciones según componentes de la misma como la forma, el contenido y el funcionamiento.

El presente artículo se organiza de la siguiente forma: la Sección II se presentan los trabajos relacionados, los cuales pretenden dar contexto a la necesidad de establecer la taxonomía; por su parte, la metodología propuesta se detalla en la Sección III; en la Sección IV se desarrollan cada una de las fases que componen la elaboración de la terminología, es decir, el establecimiento de las designaciones, las definiciones de los conceptos derivados y la definición del concepto principal de explicación, en las Secciones IV-A, IV-B, IV-C, respectivamente; en la Sección V se recoge el proceso de establecer relaciones jerárquicas entre los diferentes conceptos identificados. En la Sección VI se representa la taxonomía propuesta mediante SKOS. En la Sección VII se detalla un ejemplo de aplicación de la taxonomía sobre explicaciones reales. Por último, en la Sección VIII se presentan las

conclusiones obtenidas y las aportaciones realizadas tras la investigación.

II. Trabajos relacionados

El presente artículo nace para subsanar la carencia de un consenso a la hora de definir lo que es una explicación. Esta carencia supone una de las conclusiones alcanzadas en un trabajo previo. En [6] se realiza una revisión de literatura sistemática sobre las técnicas para generar explicaciones más utilizadas. Tras el análisis de la literatura y de otras revisiones sistemáticas se alcanza la conclusión de que es necesaria la creación de una taxonomía que permita definir el concepto de explicación.

A través del análisis de otros trabajos se puede percibir cómo diferentes autores han interpretado o concebido el término "explicación" de forma diferente. Por ejemplo, los autores de [7] afirman, basándose en el trabajo de Van Fraassen [8], que una explicación es una respuesta a una pregunta del tipo "**por qué**", en dicha respuesta una opción suele compararse con otra:

- **Pregunta**: ¿Por qué has elegido la ruta A?
- **Respuesta**: Porque la ruta A es más corta que la ruta B.

Esta conceptualización de la explicación cómo respuesta a una pregunta ha evolucionado a lo largo de los años, llegando a la inclusión de una variedad más amplia de preguntas: **por qué**, **cómo**, **quién**, **qué**, etc.

Por otro lado, en el mismo estudio y referenciando el trabajo de Achinstein en [9], se propone que una explicación debe implicar un intercambio de conocimiento entre dos actores. Esta aproximación implica que responder a una pregunta por sí solo no es suficiente, sino que la respuesta generada debe provocar una adquisición de conocimiento por parte del receptor de la misma. Esta forma de plantear las explicaciones ha conducido al uso de *managers* de diálogo o sistemas conversacionales para construir sistemas de explicabilidad, tal y como se ejemplifica en [10] y [11].

Los estudios citados en esta sección reflejan la mencionada variabilidad en el concepto de explicación, presente a lo largo de años de investigaciones en el ámbito. Este hecho permite subrayar la necesidad de establecer unas definiciones acotadas para un ámbito en concreto, ya que es posible que los conceptos de explicación varíen en función del área que se este estudiando.

III. Metodología

A la hora de desarrollar la propuesta de taxonomía se ha establecido una metodología basada en distintos métodos, normas y manuales. En específico se ha empleado cómo referencia para construir la metodología empleada la literatura que se enumera a continuación:

1. Método propuesto por Nickerson para el desarrollo de metodologías en el ámbito de los sistemas de información [12]
2. El estándar ISO 704:2022 que recoge principios y métodos para la elaboración de terminologías y sistemas de conceptos [13]
3. *Simple Knowledge Organization System* (SKOS), una serie de especificaciones y estándares para el uso de sistemas de organización del conocimiento (tesauros, esquemas de clasificación, taxonomías...) dentro de la *web* semántica [14]

A partir de las guías recogidas en la lista anterior se han establecido las diferentes fases que componen el proceso de desarrollo de la taxonomía. Dichas fases son:

1. Propósito de la taxonomía y condiciones de finalización
2. Establecimiento de la terminología
   a) Designación de conceptos e identificación de características
   b) Formalización de las definiciones de conceptos
   c) Definición del concepto principal
3. Establecimiento de relaciones
4. Representación de la taxonomía mediante SKOS

El enfoque utilizado para construir la taxonomía se basa, por lo tanto, en un proceso iterativo de identificación y definición de conceptos, según el método propuesto por Nickerson et al. Este proceso iterativo se realiza hasta cumplir una serie de condiciones de finalización preestablecidas.

Es necesario definir las condiciones de finalización de dicho proceso iterativo. Dichas condiciones se dividen en dos grupos. Por un lado se seleccionan las condiciones objetivas de entre todas las propuestas en [12]. Después se definen las condiciones subjetivas que buscan una mayor adecuación con ámbito de estudio concreto. A continuación se listan tanto las condiciones de finalización objetivas, cómo las subjetivas:

- **Condiciones de finalización objetivas**
  - *"No new dimensions or characteristics were added in the last interaction"*
  - *"Every dimension is unique and not repeated"*
  - *"Every characteristic is unique within its dimension"*
- **Condiciones de finalización subjetivas:** Concisa, robusta, comprensiva, extendible, explicativa

IV. Establecimiento de la terminología

A través del método empleado se han determinado los distintos conceptos y dimensiones que compondrán la taxonomía. Siendo el término **"explicación"** el concepto principal. Una vez identificado estos conceptos se ha elaborado una terminología utilizando para ello la norma **ISO 704:2022** como guía de referencia. La elaboración de la terminología aborda dos procesos: la designación de los conceptos y la identificación de sus características, y la definición de dichos conceptos.

IV-A. Designaciones

Según la ISO 704:2022 los conceptos que conforman una terminología se corresponden con objetos (percibidos o concebidos) que son representados mediante **designaciones** y descritos mediante **definiciones**. También es necesario tener en cuenta que a la hora de conceptualizar un objeto es necesario primero identificar sus propiedades o características.

Los conceptos se presentan en formato tabla junto con sus correspondientes designaciones y las características identificadas. Las tablas se estructuran de la siguiente forma: La columna "**Concepto**" describe la concepción abstracta del objeto; la columna "**Designación**" indica el término utilizado para representar el concepto; y la columna "**Características**" reúne las propiedades necesarias para definir el concepto.

En resumen, se han identificado cinco conceptos que se corresponden con las diferentes partes de una explicación (*Tabla I*): formato, tipo, temporalidad, autonomía y nivel de detalle. También se han identificado los conceptos derivados a cada uno de estos conceptos (*Tablas II, III, IV, V*).

Tabla I
DESIGNACIÓN Y CARACTERÍSTICAS DE LOS CONCEPTOS IDENTIFICADOS. EN NEGRITA SE SEÑALAN LOS CONCEPTOS CORRESPONDIENTES A LAS DIMENSIONES DE LA EXPLICACIÓN.

Designación	Concepto	Características
Explicación	Concepto principal. Artefacto de información proporcionado al usuario	-Entendible por un humano -Formado por datos interpretables -Información proporcionada al usuario -Mejora el entendimiento del usuario acerca del comportamiento del robot
Formato	Forma y estructura del output del sistema de explicabilidad	-Estructura de información -Representa la explicación -Propiedad para clasificar la forma
Tipo	Naturaleza de la explicación	-Estrategia estructurada -Presenta la explicación al usuario -Difiere según el enfoque
Temporalidad	Momento de aplicación de la explicación	-Espacio de tiempo en el cual se proporciona la explicación
Nivel de detalle	Adaptación de la explicación al receptor de la misma	-Dependiente del usuario -Parte opcional y variable de la explicación -Unidad de medida -Relacionado con la complejidad y cantidad de la información
Autonomía	Controla cuándo se proporciona la explicación	-Mecanismo lógico -Condiciones necesarias para generar una explicación

IV-B. Formalización de las definiciones

En esta sección se establecen las definiciones de los conceptos genéricos y específicos que corresponden a las dimensiones de la explicación y que componen la taxonomía. El objetivo de estas definiciones es establecer una referencia terminológica que permita favorecer el entendimiento y la comunicación en el ámbito de la explicabilidad en robótica.

En primer lugar se ha establecido el dominio principal de la taxonomía: **eXplainable Autonomous Robot**. La definición del dominio junto con las características identificadas en la sección anterior permite construir definiciones concretas y eliminar la posible ambigüedad entre los conceptos.

Formato: El formato de la explicación indica la forma con la que los datos que componen una explicación van a presentarse al usuario final, es decir, el tipo de dato elegido para representar la información. Se puede definir de la siguiente manera:

> **Formato:** estructura de información que representa la explicación.

El formato de una explicación puede ser de tres tipos principales, en función de cómo se comunique la información: comunicación verbal, comunicación no verbal o multimodal. Dentro del formato no verbal se pueden encontrar el formato visual, el formato numérico y el formato basado en expresividad.

Tabla II
DESIGNACIÓN Y CARACTERÍSTICAS DE LOS CONCEPTOS ESPECÍFICOS DERIVADOS DE FORMATO.

Designación	Concepto	Características
Comunicación no verbal	Tipo de formato que no utiliza palabras	-Formato de la información -No utiliza palabras
Visual	Uso de imágenes y videos	-Tipo de formato no verbal -Emplea elementos multimedia que el usuario capta a través de los ojos
Numérica	Uso de números	-Tipo de formato no verbal -Emplea números
Expresividad	Uso de gesticulación y movimiento	-Tipo de formato no verbal -Comunicación a través de gestos -Uso de capacidades físicas del robot
Comunicación verbal	Tipo de formato que utiliza palabras	-Formato de la información -Utiliza palabras
Multimodal	Combinación de formatos	-Combina varios tipos de formato

- **Comunicación no verbal**: formato de explicación que no utiliza palabras para representar la información.
  - **Visual**: utiliza imágenes o video para representar la información.
  - **Numérico**: utiliza valores numerales para representar la información.
  - **Expresividad**: utiliza la capacidad física del sistema para representar la información.
- **Comunicación verbal**: formato de explicación que utiliza palabras para representar la información.
- **Multimodal**: formato de explicación que combina la comunicación verbal y no verbal.

Tipo: El tipo de una explicación refleja el contenido y la naturaleza de la misma, es decir, la estrategia que sigue el sistema para explicar el comportamiento. La definición de tipo se ha establecido de la siguiente forma:

> **Tipo**: estrategia estructurada que establece cómo se presenta la explicación al usuario. Difiere según el enfoque.

A partir del concepto genérico de tipo surgen dos conceptos específicos: el tipo de explicación centrado en el usuario y el tipo de explicación centrado en el contexto. Por su parte, las explicaciones centradas en el usuario, pueden ser a su vez de tipo probabilístico o de tipo contrastivo, mientras que las explicaciones centradas en el contexto pueden ser secuenciales o causales.

- **Centrado en el usuario**: tipo de explicación que tiene en cuenta la intuición del usuario para presentar la explicación.
  - **Probabilístico**: presenta la explicación a través de una probabilidad de éxito.

Tabla III
DESIGNACIÓN Y CARACTERÍSTICAS DE LOS CONCEPTOS ESPECÍFICOS DERIVADOS DE TIPO.

Designación	Concepto	Características
Centrado en el usuario	Prioriza la perspectiva y la intuición del usuario a la hora de presentar la explicación	-Contenido de la explicación -Orienta el funcionamiento de la explicación -Depende del usuario -Refleja un un contraste entre la intuición del usuario y la realidad
Probabilística	El contenido de la explicación se corresponde con probabilidades	-El contenido de la información se corresponde a una probabilidad -Aborda la diferencia entre la intuición del usuario y la realidad a partir del análisis de la probabilidad de éxito
Contrastiva	El contenido de la explicación se basa en la comparación entre dos opciones	-El contenido de la información se corresponde a una comparación -Aborda la diferencia entre la intuición del usuario y la realidad señalando las diferencias entre dos opciones
Centrado en el contexto	Prioriza el contexto de lo sucedido por encima de la experiencia del usuario	-Contenido de la explicación -Orienta el funcionamiento de la explicación -No depende del usuario -Refleja lo sucedido durante el comportamiento del robot
Secuencial	El contenido de la explicación es una secuencia de sucesos	-El contenido de la información se corresponde a una narración de sucesos -Busca ofrecer una explicación a través de lo ocurrido durante el comportamiento
Causal	El contenido de la explicación es el origen de un suceso	-El contenido de la información se corresponde con la causa de un suceso -Busca ofrecer una explicación a través del análisis del origen de un suceso específico

Tabla IV
DESIGNACIÓN Y CARACTERÍSTICAS DE LOS CONCEPTOS ESPECÍFICOS DERIVADOS DE TEMPORALIDAD.

Designación	Concepto	Características
Precomportamiento	Tiene lugar antes de la ejecución	-Momento establecido en el tiempo -Previo a la ejecución
Tiempo real	Tiene lugar durante la ejecución	-Momento establecido en el tiempo -Tiene lugar mientras la ejecución se está llevando a cabo
Postcomportamiento	Tiene lugar después de la ejecución	-Momento establecido en el tiempo -Tiene lugar después de la ejecución

Tabla V
DESIGNACIÓN Y CARACTERÍSTICAS DE LOS CONCEPTOS ESPECÍFICOS DERIVADOS DE AUTONOMÍA.

Designación	Concepto	Características
Reactiva	El disparador de la explicación es una solicitud	-Se produce una solicitud de explicación
Proactiva	El disparador de la explicación depende de ciertos eventos producidos durante la ejecución	-No depende del usuario -La explicación se proporciona tras ocurrir ciertos eventos predefinidos

  - **Contrastivo**: presenta la explicación a partir del contraste entre dos opciones.
- **Centrado en el contexto**: utiliza el contexto para presentar la información.
  - **Secuencial**: presenta la explicación mediante una secuencia de sucesos.
  - **Causal**: presenta la explicación a partir del origen de un suceso.

Temporalidad: Toda explicación se debe producir en un momento en el tiempo, esta característica es representada por el concepto de "**Temporalidad**", el cual se define de la siguiente forma:

> **Temporalidad**: espacio de tiempo en el cual se proporciona la explicación.

Partiendo de esta definición se pueden derivar tres conceptos diferentes correspondientes a los tres momentos en los que se puede ofrecer la explicación: antes del comportamiento, durante el comportamiento y después del comportamiento.

- **Precomportamiento**: la explicación se proporciona antes del inicio del comportamiento del sistema.
- **Tiempo real**: la explicación se proporciona mientras el sistema está realizando su comportamiento.
- **Postcomportamiento**: la explicación se proporciona una vez el comportamiento ha finalizado.

Autonomia: Un sistema de explicabilidad debe decidir cuando proporciona una explicación sobre un comportamiento concreto. El concepto que abarca esta decisión se ha designado cómo "**Autonomía**".

> **Autonomía**: Grado de independencia del mecanismo que provee la explicación.

Además, dicha decisión puede tomarse de forma proactiva o de forma reactiva.

- **Proactiva**: el sistema decide cuando se proporciona la explicación.
- **Reactiva**: la explicación es proporcionada bajo demanda.

Nivel de detalle: Una característica importante de la explicabilidad es la capacidad del sistema de adaptarse a diferentes tipos de usuario según los conocimientos o el contexto de los mismos. Esta característica se ha bautizado cómo el nivel de detalle de la explicación, y se define de la siguiente manera:

Nivel de detalle: estrategia estructurada que establece cómo se presenta la explicación al usuario. Difiere según el enfoque.

IV-C. Definición del concepto principal

El concepto de "**explicación**" toma el papel en la presente taxonomía de concepto principal, siendo el término inicial del que derivan el resto, además de ser el primer nivel de la jerarquía de conceptos. Este concepto surge de la necesidad de dar respuesta a la pregunta "*¿qué es una explicación?*" dentro del contexto de XAR.

Antes de establecer la definición del concepto es conveniente realizar un análisis de las definiciones proporcionadas a lo largo de la literatura. Existe una tendencia a considerar que una explicación debe ser **entendible**, **transparente** e **interpretable**. Sin embargo, en este trabajo solamente se ha considerado la entendibilidad cómo una característica inherente a la explicación.

La transparencia y la interpretabilidad no pueden formar parte de la definición de explicación. La razón es que la explicación no tiene cómo requisito ser transparente sino que dota de transparencia al proceso que se desea explicar. Por otro lado, la interpretabilidad no debe ser una característica de la explicación sino de los datos o información que se utilizan para generarla. El agente receptor de la explicación no debe interpretarla sino entenderla. Mientras que el agente encargado de generar una explicación debe ser capaz de interpretar los datos para inferir la explicación correspondiente.

Por otro lado, una de las tendencias más populares en la literatura es la de plantear el concepto de explicación cómo una respuesta a una pregunta, es decir, un intercambio de comunicación entre el humano (el cual hace la pregunta) y el sistema (el cual responde). Este método de representar la explicación basado en un sistema de pregunta-respuesta en realidad gestiona cuándo el sistema debería proporcionar la explicación, lo cual se corresponde con el concepto de **autonomía**.

Junto a este concepto de autonomía, el presente trabajo ha identificado otros cuatro conceptos más que forman parte de la explicación: **formato**, **tipo**, **temporalidad** y **nivel de detalle**. Se ha considerado que la definición de la explicación debería ser independiente de cada una de estas partes. Por esta razón, y tras analizar las características identificadas y reflejadas a lo largo de la sección anterior se propone la siguiente definición de explicación:

Explicación: información proporcionada al usuario y entendible por el mismo, que esta formada por datos interpretables, y mejora la comprensión acerca del comportamiento del sistema.

V. Establecimiento de relaciones

En esta sección se definen las relaciones jerárquicas establecidas entre los conceptos identificados. En la taxonomía establecida hay dos tipos de relaciones según la ISO 704:2022:

- **Relaciones partitivas**: Son aquellas en las que un concepto concreto es considerado parte de otro concepto con el que está relacionado.

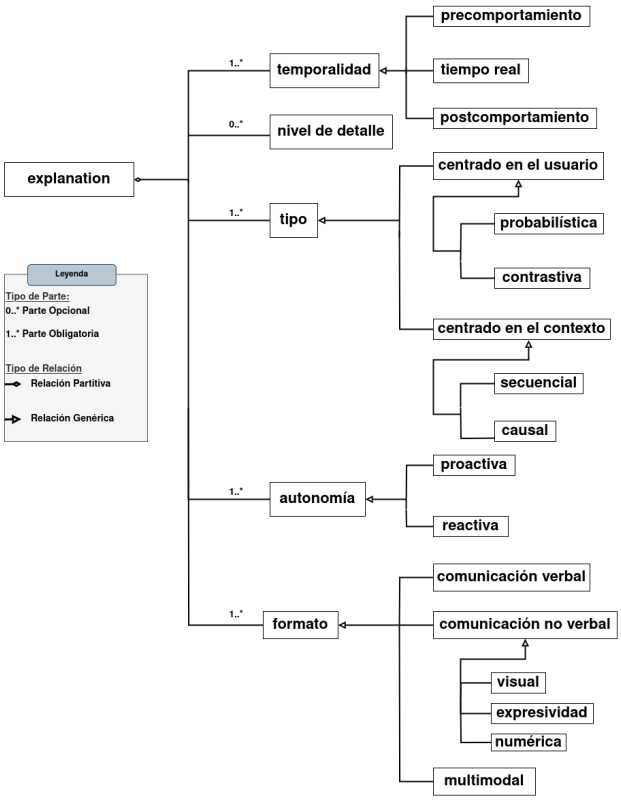


Figura 1. Diagrama UML de conceptos.

- **Relaciones genéricas**: Son aquellas en las que un concepto concreto es considerado un tipo o variante de otro concepto con el que está relacionado.

Tal y cómo ya se ha mencionado el concepto "**explicación**" es el concepto principal y tiene una relación partitiva con los cinco conceptos del siguiente nivel de la jerarquía. En las relaciones partitivas, dichas partes pueden ser de carácter obligatorio o no. De las cinco partes identificadas solamente el nivel de detalle es una parte opcional ya que depende del criterio con el que se ha generado la explicación, es decir, en un sistema de generación de explicaciones el nivel de detalle depende de la visión del desarrollador del sistema, pudiendo definir distintos niveles de detalle en función de su objetivo. Esta característica provoca que el concepto de "nivel de detalle", a su vez, no tenga conceptos específicos relacionados.

Por otro lado, todas las relaciones derivadas de cada una de estas partes se han considerado relaciones genéricas.

Todas las relaciones definidas se ilustran a través del diagrama UML de conceptos, ilustrado en la *Fig. 1*. En el diagrama se recoge el árbol jerárquico de la taxonomía. Se ilustran las relaciones partitivas y genéricas así cómo la obligatoriedad de las partes del concepto "**Explicación**".

VI. Representación de la taxonomía mendiante SKOS

En esta sección se detalla la modelización de la taxonomía mediante SKOS, lo cual permite representar los conceptos de una forma compatible con *Resource Description Framework*

(RDF). Para realizar esta representación se ha hecho uso de la sintaxis **TURTLE** [15] para RDF [16], generando así un documento *.ttl*.

SKOS utiliza etiquetas para definir los conceptos. De esta forma, *skos:prefLabel* establece la designación del concepto y *skos:definition* la definición del mismo. Se ha utilizado OWL (Web Ontology Language) [17], en concreto la etiqueta *owl:hasPart*, para definir las relaciones partitivas. En el *Listado 1* se muestra un ejemplo del código generado en el que se recoge la definición del concepto de "**Explicación**" en SKOS.

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#>
    .
@prefix ex: <http://robotics.org/explainability#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .


ex:explanation rdf:type skos:Concept;
        skos:prefLabel "Explicacion"@es;
        skos:definition "Informacion proporcionada
    al usuario y entendible por el mismo, que esta
    formada por datos interpretables, y mejora el
    entendimiento acerca del comportamiento del
    sistema."@es;
        skos:narrower ex:temporality, ex:
    level_of_detail, ex:format, ex:type, ex:
    autonomy;
        owl:hasPart ex:temporality, ex:
    nivel_de_detalle, ex:format, ex:type, ex:
    autonomy .
```

Listing 1. Concepto Explicación en SKOS

SKOS también permite definir las relaciones jerárquica entre los conceptos, para ello se hace uso de la etiqueta *skos:broader*, la cual indica que el sujeto es un concepto más específico (es decir, el hijo) de un concepto genérico. Por otro lado, la etiqueta *skos:narrower* indica la relación contraria en la que el sujeto es un concepto genérico (el padre) y está relacionado con conceptos específicos. En el *Listado 2* se muestra un ejemplo de la definición de este tipo de relaciones jerárquicas. En dicho ejemplo se pueden ver los conceptos específicos de temporalidad (precomportamiento, tiempo real y postcomportamiento).

```

ex:prebehavior rdf:type skos:Concept;
        skos:prefLabel "Precomportamiento"@es;
        skos:definition "La explicacion se
    proporciona antes del inicio del comportamiento
     del sistema."@es;
        skos:broader ex:temporality .

ex:real_time rdf:type skos:Concept;
        skos:prefLabel "Tiempo Real"@es;
        skos:definition "La explicacion se
    proporciona mientras el sistema esta realizando
     su comportamiento."@es;
        skos:broader ex:temporality .

ex:postbehavior rdf:type skos:Concept;
        skos:prefLabel "Postcomportamiento"@es;
        skos:definition "La explicacion se
    proporciona una vez el comportamiento ha
    finalizado."@es;
        skos:broader ex:temporality .
```

Listing 2. Conceptos específicos de temporalidad

Para asegurar que el código SKOS desarrollado es válido y legible por máquinas y aplicaciones es necesario validarlo. Este proceso permite también validar que la jerarquía establecida tiene coherencia y que define bien todas las relaciones. Existen

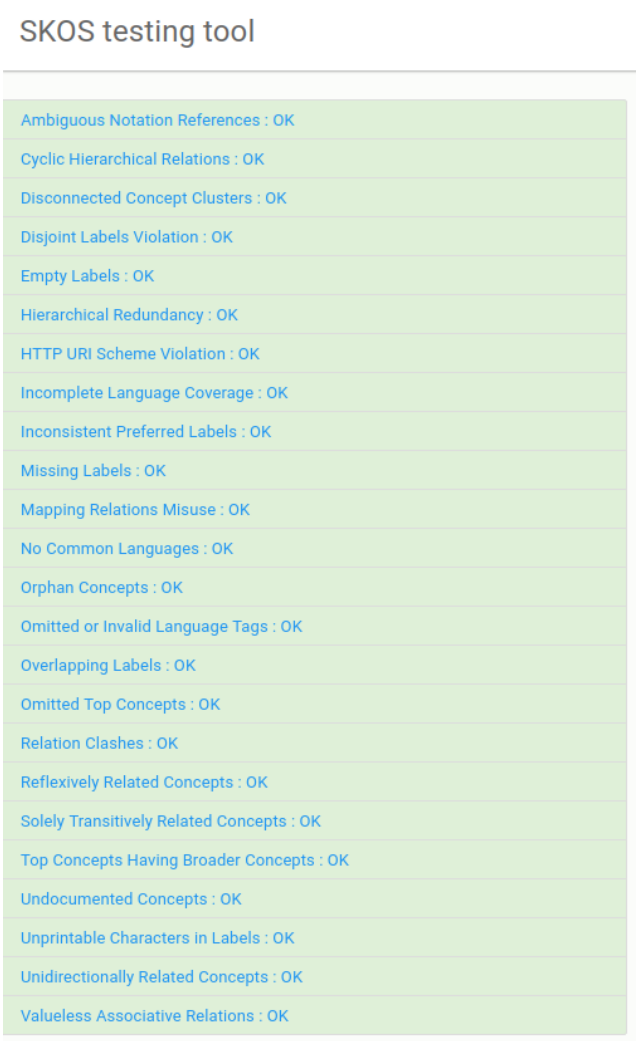


Figura 2. Tests aplicados a través de la herramienta SKOS Testing Tool.

múltiples herramientas para validar el código *ttl* desarrollado, en este caso se ha utilizado la herramienta *on-line* **SKOS Testing Tool**[1]. La *Fig. 2* recoge la lista de Test aplicados para dicha verificación.

VII. Ejemplificación del uso de la taxonomía

En esta sección se elaboran una serie de ejemplos de la aplicación de la taxonomía propuesta sobre casos reales de explicaciones en robots. Para realizar esta ejemplificación se ha partido de la revisión de literatura realizada en [18]. De entre los 22 artículos analizados en la revisión se han seleccionado 5 muestras de forma aleatoria. Empleando la taxonomía propuesta se ha realizado una clasificación de las explicaciones proporcionadas en cada una de las experimentaciones presentadas en los 5 artículos seleccionados. Los resultados se recogen en la *Tabla VI*. En dicha tabla se pueden observar 4 de las 5 partes que componen una explicación. La columna "ID" hace referencia al identificador de la muestra según la *Tabla 7* de la revisión de literatura utilizada. En la

[1]Accesible desde: https://skos-play.sparna.fr/skos-testing-tool/

columna "Formato", C.V. hace referencia a "Comunicación Verbal" y C.N.V. a "Comunicación No Verbal". Por otro lado, en la columna "Tipo", C.C hace referencia a "Centrado en el Contexto", mientras que C.U. proviene de "Centrado en el Usuario". El nivel de detalle no se ha ilustrado en la tabla porque ninguno de los cinco artículos analizados presenta un nivel de detalle definido.

Tabla VI
EJEMPLO DE LA TAXONOMÍA

ID	Temporalidad	Autonomía	Formato	Tipo
AD_3	Precomportamiento	Proactiva	C.V.	C.C.: Secuencial
AD_9	Tiempo real	Proactiva	C.V.	C.U.: Contrastiva
AD_{12}	Postcomportamiento	Reactiva	C.V.	C.C.: Causal
AD_7	Postcomportamiento	Reactivo	C.N.V.: Visual	C.C.: Secuencial
AD_6	Postcomportamiento	Reactivo	C.N.V.: Numérica	C.U.: Probabilísitca

A través del ejemplo mostrado en esta sección se refleja la capacidad de la taxonomía para descomponer, conceptualizar y clasificar las explicaciones generadas en las diferentes propuestas experimentales que pueden encontrarse en la literatura.

VIII. CONCLUSIONES

La tendencia a la automatización de tareas junto con los avances tecnológicos y la inteligencia artificial han propiciado un aumento de la presencia de sistemas robóticos autónomos en la sociedad. La convivencia e interacción del ser humano con estos sistemas está marcada por dos características fundamentales: la confianza y la seguridad.

Siendo estos dos conceptos clave en los entornos HRI, es evidente la necesidad de estudiar y avanzar en el campo de la explicabilidad aplicada a este tipo de sistemas, pues esta disciplina no solo mejora el entendimiento entre humanos y robots, y por tanto la confianza, sino que también tiene un impacto directo en la seguridad, pues desarrollar sistemas explicables facilita la detección, la comprensión y la mitigación de incidencias e, incluso, amenazas.

En el presente artículo se ha abordado una de las problemáticas principales de XAR, la cual consiste en la ausencia de definiciones consensuadas sobre en qué consiste una explicación. Establecer definiciones de conceptos es un paso imprescindible a la hora de abordar un campo tan complejo y asbtracto cómo es el XAR. La investigación realizada y recogida en este artículo ha generado las siguientes contribuciones en el mencionado área:

- Una terminología del concepto explicación y las partes que lo componen, junto con sus definiciones.
- Una taxonomía jerárquica que recoge los conceptos identificados y cómo se relacionan entre ellos.
- La representación de dicha taxonomía en un formato compatible con RDF.

Por otro lado, como trabajo futuro se plantea desarrollar y aplicar un método para validar la utilidad real de la taxonomía. El procedimiento de validación consistirá en contar con la participación de expertos en robótica y explicabilidad en IA. Se seleccionarán una serie de casos en los que se generen explicaciones en sistemas robóticos. Dichos casos se presentarán a los expertos junto a la taxonomía propuesta y se les pedirá que evalúen mediante un cuestionario en qué medida cada componente de la taxonomía está presente en las explicaciones analizadas. Además deberán evaluar si la taxonomía permite una descomposición clara de las partes de la explicación y si existen elementos redundantes o faltantes. Este procedimiento proporcionará unos resultados que serán analizados de manera cualitativa y cuantitativa con el objetivo de validar la taxonomía, así como de refinarla de forma iterativa.

El trabajo aquí recogido se presenta cómo una propuesta de términos y clasificación que tiene el objetivo de no solo de proporcionar un contexto conceptual a las investigaciones realizadas en XAR, sino también de mejorar la propia taxonomía a través de la retroalimentación de la comunidad investigadora.

AGRADECIMIENTOS

Esta investigación resulta del Proyecto Estratégico TESCAC (C82/23), fruto del convenio de colaboración suscrito entre el Instituto Nacional de Ciberseguridad (INCIBE) y la Universidad de León. Esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation). Esta investigación también ha sido financiada por el proyecto EDMAR: Grant PID2021-126592OB-C21 financiado por MCIN/AEI/10.13039/501100011033 y por FEDER, UE. David Sobrín Hidalgo agradece a la Universidad de León la beca proporcionada para financiar sus estudios de doctorado.

REFERENCIAS

[1] T. B. Sheridan, "Human–robot interaction: status and challenges," *Human factors*, vol. 58, no. 4, pp. 525–532, 2016.

[2] G. Papagni and S. Koeszegi, "Understandable and trustworthy explainable robots: A sensemaking perspective," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 13–30, 2020.

[3] T. Sakai and T. Nagai, "Explainable autonomous robots: a survey and perspective," *Advanced Robotics*, vol. 36, no. 5-6, pp. 219–238, 2022. [Online]. Available: https://doi.org/10.1080/01691864.2022.2029720

[4] A. Roque and S. K. Damodaran, "Explainable ai for security of human-interactive robots," *International Journal of Human–Computer Interaction*, vol. 38, no. 18-20, pp. 1789–1807, 2022.

[5] V. Matellán, F.-J. Rodríguez-Lera, Á.-M. Guerrero-Higueras, F.-M. Rico, and J. Ginés, "The role of cybersecurity and hpc in the explainability of autonomous robots behavior," in *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 2021, pp. 1–5.

[6] D. S. Hidalgo, I. G. Fernández, M. Á. G. Santamarta, A. C. Vega, Á. M. G. Higueras, F. J. R. Lera, and V. M. Olivera, "Explicabilidad aplicada a ciberseguridad en sistemas autónomos: Un enfoque basado en llms para la detección forense de amenazas," in *IX Jornadas Nacionales de Investigación En Ciberseguridad*. Antonia M. Reina Quintero, 2024, pp. 644–651.

[7] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 33, pp. 673–705, 2019.

[8] B. Van Fraassen, "Empiricism in the philosophy of science." in *Images of science: Essays on realism and empiricism*. University of Chicago Press, 1985, pp. 245–308.

[9] P. Achinstein, *The nature of explanation*. Oxford University Press, USA, 1983.

[10] V. Maraev, E. Breitholtz, C. Howes, and J.-P. Bernardy, "Why should i turn left? towards active explainability for spoken dialogue systems." in *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, 2021, pp. 58–64.

[11] B. Galitsky and B. Galitsky, "Conversational explainability," *Artificial Intelligence for Customer Relationship Management: Keeping Customers Informed*, pp. 415–445, 2020.
[12] R. C. Nickerson, U. Varshney, and J. Muntermann, "A method for taxonomy development and its application in information systems," *European Journal of Information Systems*, vol. 22, no. 3, pp. 336–359, 2013.
[13] O. I. de Normalización, *ISO 704:2022 Terminology work - Principles and methods = Travail terminologique - Principes et méthodes*, 4th ed. Organización Internacional de Normalización, 2022.
[14] A. Miles and S. Bechhofer, "Skos simple knowledge organization system reference," *W3C recommendation*, 2009.
[15] W. W. W. Consortium *et al.*, "Rdf 1.1 turtle: Terse rdf triple language," *W3C Recommendation*, 2014.
[16] D. Beckett, T. Berners-Lee, E. Prud'hommeaux, and G. Carothers, "Rdf 1.1 turtle," *World Wide Web Consortium*, pp. 18–31, 2014.
[17] G. Antoniou and F. v. Harmelen, "Web ontology language: Owl," *Handbook on ontologies*, pp. 91–110, 2009.
[18] D. Sobrín-Hidalgo, Á. M. Guerrero-Higueras, and V. Matellán-Olivera, "Generating explanations for autonomous robots: a systematic review," *IEEE Access*, 2025.

Sesión Transferencia I

Protección de Información Biométrica con Fuzzy Vault y Cifrado Homomórfico

Ángela Barriga Rodríguez
Mobbeel Solutions SL
Cáceres, España
abarriga@mobbeel.com

Álvaro Hernández Martín
Mobbeel Solutions SL
Cáceres, España
ahernandez@mobbeel.com

Belén González Sánchez
Mobbeel Solutions SL
Cáceres, España
bgonzalez@mobbeel.com

Daniel Santos Anés
Mobbeel Solutions SL
Cáceres, España
dsantos@mobbeel.com

Javier Nieto Acero
Mobbeel Solutions SL
Cáceres, España
jnieto@mobbeel.com

Santiago García Gil
Mobbeel Solutions SL
Cáceres, España
sgarcia@mobbeel.com

***Resumen*—Los sistemas biométricos basados en Deep Learning proporcionan un alto nivel de seguridad, pero es esencial seguir investigando para añadir capas adicionales de protección y cumplir con los principios de la Norma ISO/IEC 24745. Un enfoque destacado es el uso de Referencias Biométricas Renovables, que garantizan la irreversibilidad, no interoperabilidad y revocabilidad de los datos, cumpliendo con los requisitos de dicha norma. El trabajo de investigación de Mobbeel se centra en técnicas como Fuzzy Vault y Cifrado Homomórfico para proteger las plantillas biométricas. Los resultados obtenidos demuestran que estas soluciones refuerzan la seguridad, aunque su implementación genera un coste computacional significativo, lo cual debe ser evaluado antes de su despliegue en entornos productivos.**

***Index Terms*—Biometría, Criptografía, Seguridad, Privacidad**

Tipo de contribución: *Transferencia*

I. Introducción

La biometría se ha consolidado como una tecnología clave en la verificación de identidad, proporcionando métodos más seguros y eficientes que los tradicionales, como las contraseñas o los tokens. No obstante, su adopción generalizada ha aumentado los riesgos asociados a la privacidad y seguridad de los datos biométricos. La sensibilidad inherente de esta información, junto con las vulnerabilidades frente a ataques específicos, filtraciones o usos indebidos, plantea desafíos críticos que requieren atención y medidas de protección adecuadas. [1].

En este sentido, la norma internacional ISO/IEC 24745 [2] establece una serie de requisitos fundamentales para la protección de plantillas biométricas, destacando en concreto cuatro principios clave: irreversibilidad, no interoperabilidad, confidencialidad y revocabilidad (*irreversibility*, *unlinkability*, *confidentiality*, *revocability* respectivamente). Es decir, que: (i) la reconstrucción de los datos biométricos originales desde las plantillas no sea posible, (ii) no exista una correlación entre plantillas usadas en distintos sistemas, esto es, que no sean idénticas a nivel de formato de datos y no puedan ser reutilizadas, (iii) que las plantillas estén protegidas contra accesos no autorizados y (iv) que si el sistema se ve comprometido las plantillas puedan eliminarse y sustituirse. El cumplimiento de esta norma requiere la combinación de la biometría con diferentes técnicas de seguridad, lo cual es aún un reto por resolver debido a su complejidad técnica [3].

En esta misma línea, el Centro Criptológico Nacional (CCN) ha destacado la importancia de adoptar enfoques como el uso de Referencias Biométricas Renovables (RBRs), con el fin de reforzar la seguridad y privacidad en los sistemas biométricos [4]. Esta recomendación busca reducir el impacto ante posibles brechas de seguridad mediante plantillas biométricas que puedan renovarse sin comprometer la información original del usuario ni su privacidad, pero manteniendo las prestaciones del sistema biométrico (o con un impacto muy limitado en su rendimiento) [1].

En este artículo, se analizan las técnicas de seguridad estudiadas por Mobbeel para la protección de plantillas biométricas faciales obtenidas mediante aprendizaje profundo (*Deep Learning*) [5]. En particular, se presentan dos técnicas prometedoras: el esquema Fuzzy Vault (FV) [6] y el Cifrado Homomórfico (CH) [3].

II. Protección de Información Biométrica

La protección de información biométrica supone un desafío crítico debido a la naturaleza sensible de este tipo de datos, definidos como sujetos de especial protección según el Reglamento General de Protección de Datos (RGPD), Artículo 9 apartado 1 [7]. En esta sección analizamos cuáles son las principales amenazas a las que se ven expuestos los datos biométricos y qué soluciones existen en la actualidad para hacer frente a este reto de seguridad.

II-A. Principales amenazas

Los riesgos asociados a la seguridad y privacidad de los datos personales se han incrementado considerablemente debido a la adopción generalizada de sistemas biométricos. Aunque estos ofrecen altos niveles de seguridad y comodidad no están exentos de riesgos. Entre las amenazas que atañen a estos sistemas destacan especialmente:

- el robo de plantillas biométricas,
- la reutilización indebida de datos comprometidos,
- los ataques de reconstrucción inversa de las plantillas.

En los sistemas biométricos tradicionales, donde se utilizaban cálculos antropométricos estandarizados, el robo de

plantillas biométricas representaba un riesgo crítico debido a la naturaleza permanente e irreemplazable de estos datos. A diferencia de contraseñas o tokens, una vez comprometidas, las plantillas biométricas no podían ser revocadas ni sustituida fácilmente [2]. Además, el impacto se agravaba por el uso indebido de datos biométricos robados, ya que una misma plantilla podía ser explotada para acceder o registrarse en múltiples servicios y plataformas. Sin embargo, con la evolución de los sistemas biométricos basados en *Deep Learning*, estos riesgos se han mitigado, ya que las plantillas generadas no son una réplica directa del rasgo original, sino representaciones transformadas y específicas de cada sistema.

Por otro lado, los ataques de reconstrucción inversa buscan reconstruir parcial o totalmente los datos biométricos originales a partir de las plantillas sustraídas, violando directamente el principio de irreversibilidad establecido por la norma ISO/IEC 24745 [2]. Estos ataques se dividen en dos categorías: los de caja blanca, en los que el modelo del sistema biométrico es manipulado directamente por el atacante, lo que permite reconstrucciones más precisas; y los de caja negra, que se basan únicamente en consultas al sistema y la observación de sus respuestas, alcanzando reconstrucciones menos precisas. Aunque efectivos en laboratorio, ambos tipos de ataques enfrentan dificultades significativas en entornos reales [8].

Si bien los sistemas biométricos de nueva generación basados en *Deep Learning* ofrecen un alto grado de seguridad, es recomendable ir un paso más allá incorporando técnicas criptográficas avanzadas para aportar una capa extra de seguridad y aportar factores como la revocabilidad de datos.

II-B. Soluciones convencionales vs. enfoques avanzados

Tradicionalmente, las plantillas biométricas se han almacenado en bases de datos sin modificaciones tras su extracción, protegiéndolas con medidas físicas, lógicas y canales seguros. También se han empleado técnicas criptográficas tradicionales (CT) como AES o RSA para garantizar confidencialidad en almacenamiento y transmisión. Sin embargo, estos métodos requieren descifrar los datos antes de procesarlos, exponiéndolos temporalmente a ataques [9].

Por su parte, algunos esquemas de hashing alteran la estructura espacial de los vectores de características, dificultando la autenticación biométrica. Esto limita la funcionalidad del sistema y no cumple plenamente con la norma ISO/IEC 24745 en términos de irreversibilidad y desvinculación [2]. Para abordar estas limitaciones, han surgido enfoques avanzados como FV y CH.

Con FV las plantillas biométricas son protegidas combinando datos biométricos con claves criptográficas, generando puntos falsos (*chaff points*) para dificultar la ejecución de ataques [6]. Ofrece tolerancia a errores y resistencia frente a reconstrucción inversa [6], [10], pero presenta desafíos en rendimiento, especialmente en dispositivos con recursos limitados [11].

Por su parte, CH permite operar sobre datos cifrados sin necesidad de descifrarlos, garantizando confidencialidad absoluta incluso en entornos vulnerables [3]. Además, cumple plenamente con ISO/IEC 24745 y ofrece alta resistencia frente a ataques externos gracias al procesamiento seguro de los datos cifrados [3].

La Tabla I compara CT, FV y CH según los criterios de seguridad de la norma ISO, el CNN y su coste computacional.

Tabla I
COMPARATIVA ENTRE CIFRADO TRADICIONAL Y TÉCNICAS AVANZADAS EN TÉRMINOS DE SEGURIDAD

Característica	**CT**	**FV**	**CH**
Confidencialidad	Alta	Alta	Muy Alta
Irreversibilidad	Baja	Alta	Alta
Desvinculación	Baja	Alta	Alta
Renovabilidad	Baja	Alta	Media
Complejidad computacional	Baja-Media	Media-Alta	Muy Alta
Cumplimiento ISO 24745	Parcial	Alto	Completo

III. APLICACIONES DE BIOMETRÍA SEGURA

Dado el potencial presentado por FV y CH, en la siguiente sección se exploran estas técnicas en profundidad, analizando su aplicación en casos de uso reales de Mobbeel y evaluando su rendimiento y nivel de seguridad.

III-A. Fuzzy Vault para Cifrado de Plantillas Biométricas

El esquema criptográfico FV protege los datos combinándolos con claves criptográficas, impidiendo su reconstrucción directa y dificultando los ataques al añadir ruido en forma de *chaff points* [6]. Estos puntos, generados aleatoriamente, se mezclan con los datos dentro de un *vault*, complicando la recuperación de la clave secreta sin el dato original o uno altamente similar [6].

III-A1. Proceso de verificación indirecta: FV protege las plantillas biométricas almacenadas, sin desbloquearlas hasta el proceso de verificación. El dato biométrico del usuario se usa para identificar los puntos genuinos en el *vault*, descartando los *chaff points*. Si la plantilla coincide, se aplica un proceso de interpolación polinómica para regenerar la clave criptográfica original, garantizando a la vez que las plantillas no se almacenen directamente y sean resistentes a ataques de reconstrucción [12].

Una ventaja clave de FV es la verificación indirecta: la plantilla biométrica capturada no se almacena ni compara directamente, sino que se utiliza para intentar regenerar la clave original dentro del *vault*. Si esto se logra, el usuario se considerará verificado [13].

En biometría facial, este enfoque es efectivo, ya que las características extraídas por *Deep Learning* son altamente discriminativas y robustas frente a variaciones intra-clase (iluminación, expresiones), permitiendo una regeneración precisa sin comprometer la privacidad [14]–[16].

FV mejora la seguridad y privacidad al evitar el almacenamiento directo de datos biométricos, reduciendo riesgos ante robos o filtraciones. Incluso si un atacante comprometiera el *vault*, la presencia de *chaff points* y la necesidad del vector biométrico correcto hacen extremadamente difícil recuperar información sensible [6], [17].

III-A2. Ajustes de FV para el caso de uso en Mobbeel: FV se incorpora al sistema biométrico de Mobbeel utilizando la implementación en C++ de la técnica *Guruswami-Sudan*, tal como se propone en [12], mediante la librería *Thimble* [18]. En esta implementación, se crea un *vault* a partir de la plantilla biométrica generada en la etapa de modelado.

Como parte del preprocesado, se realiza una cuantización en intervalos equi-probables, seguida de una binarización

mediante código linealmente separable y un mapeo de las posiciones con valor 1 en intervalos de 4. Esta configuración se establece tras compararla con otras técnicas, evaluando cuál ofrece mayor precisión en la identificación de pares.

Para que *Guruswami-Sudan* pueda reconstruir el polinomio con éxito, es necesario que $w \geq \sqrt{u(k-1)}$, siendo:

- **w**: número de puntos genuinos entre ambas caras, es decir, los puntos en los que coincide cada plantilla biométrica.
- **u**: número de puntos de la cara para desbloquear el *vault*.
- **k**: grado del polinomio secreto.

Evaluando las distancias entre las plantillas biométricas se concluye que se necesitan al menos 300 puntos (de los 512 que forman la plantilla biométrica que genera la tecnología de Mobbeel) en común entre pares para que se considere la misma identidad y que, por lo tanto, el *vault* pueda abrirse. Siendo w 300 y u un valor entre 510-520 (512 alterados por los *chaff points*), se procede a hacer pruebas con distintos valores de k, con una multiplicidad de 3.

III-B. Cifrado Homomórfico para Operaciones Biométricas Seguras

El CH es una técnica criptográfica avanzada que permite realizar operaciones matemáticas sobre datos cifrados sin necesidad de descifrarlos previamente. Matemáticamente, si $E(x)$ representa el cifrado de un dato x, el CH permite realizar operaciones como $E(x) \oplus E(y) = E(x + y)$ o $E(x) \otimes E(y) = E(x \times y)$, donde el resultado cifrado corresponde exactamente al cifrado de la operación sobre los datos originales. Esta propiedad fundamental hace posible que terceros procesen información sensible sin acceder a su contenido real, preservando así la confidencialidad e integridad de los datos durante todo su ciclo de vida.

Existen tres categorías principales de CH: parcialmente homomórfico (PHE), que admite un tipo limitado de operaciones; algo homomórfico (SHE), que permite un número limitado de operaciones; y totalmente homomórfico (FHE), que teóricamente permite un número ilimitado de operaciones. En aplicaciones biométricas, el FHE es particularmente relevante, aunque su implementación práctica enfrenta desafíos significativos de rendimiento [19].

III-B1. Ventajas y desafíos: Entre sus ventajas, el CH garantiza irreversibilidad, renovabilidad y no vinculabilidad de las plantillas biométricas, cumpliendo así con los requisitos esenciales para la protección de datos biométricos y alineándose con las normativas ISO y del CCN. Además, como señalan Yalavarthi et al. en [20], este enfoque es robusto frente a ataques de dispositivos robados, ataques de tasa de falsa aceptación (FAR). Por otro lado, el CH contribuye a prevenir el robo de identidad al garantizar que los datos biométricos se mantengan protegidos, incluso en caso de una brecha de seguridad [21].

En cuanto a los desafíos, el principal obstáculo para su implementación práctica es el alto coste computacional. Para que este enfoque sea viable, es necesario realizar optimizaciones significativas. Algunas implementaciones han logrado reducir el tamaño del cifrado mediante técnicas de compresión [21] y otras recurren a técnicas de aceleración hardware tanto CPU (HEXL) [22] como GPU (CUDA) [23].

Finalmente, aunque el CH proporciona mayor seguridad, los esquemas de CH pueden estar expuestos a diferentes tipos de ataques como: de decodificación, duales, de canal lateral y de recuperación de clave [21].

III-B2. Incorporación de CH para el caso de uso en Mobbeel: Para complementar la tecnología de Mobbeel con CH, evaluamos dos librerías de Python: TenSEAL [22], con aceleración HEXL, y PhantomPHE [23], con aceleración CUDA. Ambas utilizan el algoritmo Cheon-Kim-Kim-Song (CKKS), ya que es el único que soporta datos numéricos reales sin aproximarlos a enteros.

El CH requiere un contexto para crear el campo numérico que encripta y procesa los datos. En pruebas preliminares, se concluye que un contexto con polinomios de módulo entre 4096 y 8192, escala global entre 220 y 260, y coeficientes modulares entre [30, 20, 30] y [60, 40, 40, 60] ofrece un equilibrio adecuado entre seguridad y eficiencia. También observamos que, a menor contexto, los tiempos de ejecución se reducen, pero a costa de disminuir la seguridad.

Además, evaluamos el tiempo de ejecución con la aceleración proporcionada por cada librería, sin encontrar diferencias significativas entre HEXL (TenSEAL) y CUDA (PhantomPHE), posiblemente debido a que la aceleración de CH en CUDA sigue siendo un proceso experimental.

III-C. Comparativa de rendimiento: Sin cifrado, FV y CH

Para comparar el rendimiento de las alternativas examinadas (FV y CH), hemos seleccionado un subconjunto reducido del dataset Labelled Faces in the Wild (LFW) [24], seleccionando 2000 pares de gran similitud para llevar al extremo el sistema: 1000 identidades coincidentes y 1000 identidades no coincidentes. Es importante señalar que esta base de datos, de pequeñas dimensiones y utilizada específicamente para esta tarea, no es de propósito general y no permite extraer conclusiones sobre la precisión real de la tecnología biométrica de Mobbeel. Los resultados obtenidos en este trabajo no son equivalentes a los reportados en evaluaciones de referencia como NIST FRTE, donde la tecnología de Mobbeel ha demostrado desempeños significativamente superiores. No obstante, este conjunto es adecuado para comparar los distintos enfoques explorados en esta investigación.

La comparativa se realiza entre tres configuraciones biométricas:

1. **Sin cifrado:** Utiliza directamente las plantillas biométricas generadas por *Deep Learning*, sin aplicar ningún mecanismo de protección.
2. **FV (Fuzzy Vault):** Protege las plantillas biométricas aplicando el esquema *Fuzzy Vault*.
3. **CH (Cifrado Homomórfico):** Cifra las plantillas biométricas mediante cifrado homomórfico utilizando la librería *TenSEAL*.

Tabla II
COMPARATIVA DE PRECISIÓN Y TIEMPOS DE EJECUCIÓN

Configuración	Errores FP	Errores FN	Tiempos medios (ms)
Sin cifrado	2	82	250
FV	5	157	480
CH	2	82	550

Durante las pruebas, primero se registra una identidad de cada par en el sistema biométrico; posteriormente, el otro elemento del par se emplea en el proceso de verificación. La Tabla II detalla los resultados de estas verificaciones, incluyendo el número de falsos positivos (FP), falsos negativos (FN) (identidades falsas identificadas como verdaderas y viceversa) y el tiempo medio de ejecución.

Se observa que la configuración basada en CH mantiene una precisión idéntica a la del sistema sin ninguna seguridad adicional ("Sin cifrado"), aunque con un coste computacional significativamente mayor, reflejado en el aumento del tiempo de ejecución. Por otro lado, aunque el uso de FV presenta un incremento menor en los tiempos de ejecución, el número de errores de clasificación aumenta de manera significativa, lo que lo convierte en una opción menos favorable. Para mitigar esta degradación en el rendimiento, sería posible emplear parámetros más robustos para la reconstrucción del polinomio, aunque esto implicaría un incremento en la complejidad computacional.

Todos los experimentos se ejecutaron en instancias AWS c7i.xlarge de 8 vCPUs y 16 GB de RAM.

IV. Conclusiones y trabajo futuro

Los experimentos muestran que el cifrado homomórfico (CH) ofrece una mejor combinación de seguridad y precisión en comparación con Fuzzy Vault (FV), cuyo rendimiento se ve notablemente afectado debido a la complejidad del proceso de reconstrucción de la clave criptográfica. Los resultados obtenidos destacan la viabilidad de CH en aplicaciones biométricas de verificación, aunque con un margen de mejora considerable en términos de eficiencia.

Como trabajo futuro, es crucial optimizar el rendimiento computacional del cifrado homomórfico, explorando soluciones que permitan su implementación eficiente en sistemas a gran escala. Esto podría implicar la optimización de las librerías existentes y la ejecución de algoritmos en hardware especializado, como GPUs. Otras líneas de investigación incluyen la exploración de enfoques híbridos que combinen lo mejor de ambos métodos. Así, sería interesante ampliar las pruebas presentadas en este artículo, con nuevos datos, librerías de cifrado, técnicas de *Deep Learning* y entornos de ejecución.

Asimismo, con el fin de seguir estudiando el potencial de estas técnicas de cifrado, se podría afrontar desafíos como estudiar la aplicación de CH en operaciones biométricas más exigentes computacionalmente que la verificación de pares, tales como la identificación de una identidad sobre una base de datos, lo cual sigue siendo un desafío por resolver en la academia y la industria debido a su elevado coste computacional.

Referencias

[1] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 182–206, 2023.

[2] *ISO/IEC 24745: Information technology — Security techniques — Biometric information protection*, International Organization for Standardization Std. ISO/IEC 24 745, 2011.

[3] M. e. a. Gomez-Barrero, "Bloom filter vs homomorphic encryption: Which approach protects the biometric data and satisfies iso/iec 24745?" *IEEE Access*, vol. 9, pp. 123 456–123 470, 2021.

[4] Centro Criptológico Nacional, "Ccn-tec 013: Requisitos de seguridad para aplicaciones de software," 2023, acceso: marzo, 2025. [Online]. Available: https://www.ccn.cni.es/es

[5] N. Prasad, B. Rajpal, K. K. R. Mangalore, R. Shastri, and N. Pradeep, "Frontal and non-frontal face detection using deep neural networks (dnn)," *International Journal of Research in Industrial Engineering*, vol. 10, no. 1, pp. 9–21, 2021.

[6] K. Nandakumar, A. Nagar, and A. K. Jain, "Biometric encryption based on a fuzzy vault scheme with a fast chaff generation algorithm," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 563–574, 2013.

[7] E. Parliament and C. of the European Union, "Article 9: Processing of special categories of personal data," General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, 2016. [Online]. Available: https://gdpr-info.eu/art-9-gdpr/

[8] H. Otroshi Shahreza and S. Marcel, "Template inversion attack using synthetic face images against real face recognition systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 2024.

[9] P. Huang, G. Liao, and J. Ren, "The application of aes-sm2 hybrid encryption algorithm in big data security and privacy protection," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 6, pp. 989–994, 2024.

[10] A. M. Elbanaa, R. I. Abdelfatah, and M. E. Nasr, "Securing body area networks with fingerprint cryptography and fuzzy extractor," *Journal of Engineering Research*, vol. 8, no. 3, pp. 1–16, 2024.

[11] U. Uludag, S. Pankanti, and A. K. Jain, "Fuzzy vault for fingerprints," in *Audio- and Video-Based Biometric Person Authentication*, T. Kanade, A. Jain, and N. K. Ratha, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 310–319.

[12] C. Rathgeb, J. Merkle, J. Scholz, B. Tams, and V. Busch, "Deep face fuzzy vault: Implementation and performance," *Computer Science, Computer Vision and Pattern Recognition*, November 2021.

[13] Y. H. Lee, M. Khalil-Hani, and M. N. Marsono, "Hardware-based biometric encryption implementation with gauss-jordan algorithm accelerator core in field programmable gate arrays," *IET Information Security*, vol. 5, no. 4, pp. 234–243, 2011.

[14] K. M. Alheeti, A. Gruebler, K. D. McDonald-Maier, and A. Fernando, "A fuzzy vault development based on iris images," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 9, no. 4, pp. 232–239, 2021.

[15] J. Hartloff, M. Bileschi, S. Tulyakov, J. Dobler, A. Rudra, and V. Govindaraju, "Security analysis for fingerprint fuzzy vaults," in *Biometric and Surveillance Technology for Human and Activity Identification X*, I. Kakadiaris, W. J. Scheirer, and L. G. Hassebrook, Eds., vol. 8712, International Society for Optics and Photonics. SPIE, 2013, p. 871204. [Online]. Available: https://doi.org/10.1117/12.2016635

[16] A. Kaur and P. Khanna, "Alignment-free cancelable biometric: A contemporary survey, opportunities and challenges," *IEEE Access*, vol. 8, pp. 225 445–225 470, 2020.

[17] R. Mehmood and A. Selwal, "Polynomial based fuzzy vault technique for template security in fingerprint biometrics," *International Arab Journal of Information Technology*, vol. 17, no. 6, pp. 926–934, 2020.

[18] B. Tams, P. Mihailescu, and A. Munk, "Security considerations in minutiae-based fuzzy vaults," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 985–998, May 2015.

[19] S. M. Arman, T. Yang, S. Shahed, A. Al Mazroa, A. Attiah, and L. Mohaisen, "A comprehensive survey for privacy-preserving biometrics: Recent approaches, challenges, and future directions," *Computers, Materials and Continua*, vol. 78, no. 2, pp. 2087–2110, 2024.

[20] B. Yalavarthi, A. R. Kaushik, A. Ross, V. Boddeti, and N. Ratha, "Enhancing privacy in face analytics using fully homomorphic encryption," *arXiv preprint*, April 2024.

[21] G. Pradel and C. J. Mitchell, "Privacy-preserving biometric matching using homomorphic encryption," *CoRR*, vol. abs/2111.12372, 2021. [Online]. Available: https://arxiv.org/abs/2111.12372

[22] A. Benaissa, B. Retiat, B. Cebere, and A. E. Belfedhal, "Tenseal: A library for encrypted tensor operations using homomorphic encryption," *arXiv preprint*, April 2021.

[23] H. Yang, S. Shen, W. Dai, L. Zhou, Z. Liu, and Y. Zhao, "Phantom: A cuda-accelerated word-wise homomorphic encryption library," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4895–4906, 2024.

[24] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Technical Report 07-49, October 2007.

Dynamic Industrial Cyber risk Modelling based on Evidence (DICYME)

Alberto Fernández-Isabel
Rey Juan Carlos University
Móstoles, Spain
alberto.fernandez.isabel@urjc.es

Isaac Martín de Diego
Rey Juan Carlos University
Móstoles, Spain
isaac.martin@urjc.es

Emilio L. Cano
Rey Juan Carlos University
Móstoles, Spain
emilio.lopez@urjc.es

Rubén R. Fernández
Rey Juan Carlos University
Móstoles, Spain
ruben.rodriguez@urjc.es

Javier García-Ochoa
Rey Juan Carlos University
Móstoles, Spain
javier.garciaochoa@urjc.es

Romy R. Ravines
DeNexus Inc.
Boston, United States
rr@denexus.io

Ovidio López
DeNexus Inc.
Boston, United States
ol@denexus.io

Jaume Puigbó
DeNexus Inc.
Boston, United States
jp@denexus.io

***Abstract*—The increasing prevalence of cyber threats targeting industrial infrastructures highlights the urgent need for advanced cyber risk quantification models tailored to these environments. The DICYME project, led by DeNexus in collaboration with Rey Juan Carlos University (URJC), focuses on creating valuable datasets and cybersecurity risk indicators that enhance cyber risk quantification processes, whether within our proposed framework or other cyber risk quantification models. Our contribution includes a cyber risk quantification approach that integrates curated datasets through automated data extraction, developing cybersecurity risk indicators, and statistical simulations, to enable dynamic data-driven risk assessments. These components are consolidated into a visualization tool, allowing industrial operators and insurers to analyze, quantify, and mitigate cyber risks with structured and evidence-based insights.**

***Index Terms*—Cyber risk, Decision-Making System, Company features, Impact, Threat actors**

Type of contribution: *R&D projects carried out with companies or end users of technology (limit 4 pages)*

I. Introduction

Cyberattacks targeting industrial systems, such as the Stuxnet incident in 2010 [1], have demonstrated the critical vulnerabilities of operation technology (OT) environments. These systems, essential for maintaining vital societal functions, face unique risks not addressed by traditional information technology (IT) cybersecurity frameworks. Although existing approaches to industrial cybersecurity have advanced defensive strategies [2], significant gaps remain in cyber risk modeling and quantification, particularly in OT and critical infrastructure environments [3], [4].

The Dynamic Industrial Cyber risk Modelling based on Evidence (DICYME) project has developed an innovative approach to cyber risk modeling, including the generation of curated datasets by using public and private data sources, the development of cyber security indicators with novel equations and data combination, and the visualization of data in a structured and valuable way for decision-makers. All these results are valuable outcomes for DeRisk, a cybersecurity risk assessment platform developed by DeNexus which works with well-known frameworks such as NIST cybersecurity framework (CSF) or ISO 27001.

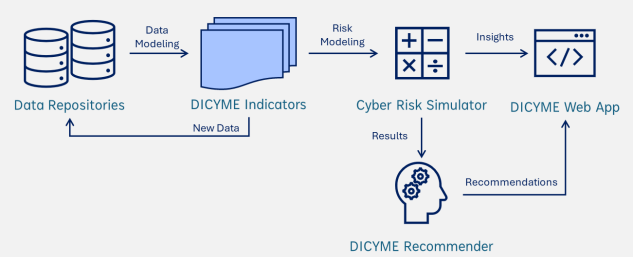


Figure 1. DICYME complete system approach.

In addition to these advancements, DICYME has developed a comprehensive system that incorporates a cyber risk quantification (CRQ) framework, combining real-time data from the aforementioned outcomes to generate quantitative cyber risk assessments through probabilistic modeling.

The rest of this paper is organized as follows. Section II overviews the workflow followed to build useful models for the quantification of industrial cyber risk. Section III presents the project's main achievements, including the complete system that groups and uses partial results to quantify an entity's cyber risk. Finally, Section IV establishes the current state of the project, detailing the ongoing work, and concludes with an explanation of the research transfer to not only DeRisk but also to third parties.

II. System workflow

DICYME follows a multi-phase workflow to produce valuable cyber risk outputs, beginning with data collection into data repositories, followed by its transformation into useful indicators and models for cyber risk quantification, and employing visualization techniques and recommendations for decision support. Finally, these outputs can be separately introduced into the DeRisk system or into a specific end-to-end system, developed within DICYME, which includes a cyber risk quantification framework as well as a recommender system currently under development (see Figure 1).

A. Evidence collection

The first step focuses on the automated and dynamic extraction of cyber risk evidence from internal and external sources,

ensuring that the data reflects the current context and allows for the analysis of changes over time. Internal sources include specific cyber-physical systems (CPS) protection platforms and network visibility solutions, while external sources encompass open source intelligence (OSINT) tools and threat intelligence data. The extracted data are normalized and transformed into formats compatible with risk modeling algorithms.

B. Probabilistic and impact modeling

Using the evidence gathered, DICYME develops indicators and models to estimate:

- Probability of incidents: Factors such as threat actor capabilities, infrastructure visibility and firmographic profiles are considered.
- Impact of incidents: Technical and financial impacts are modeled using firmographic data again, historic trends and threat intelligence.

These models use statistical techniques, such as Monte Carlo simulations, to generate dynamic risk metrics [5], [6], [7]. Additionally, statistical distributions like Poisson are used to model the occurrence of events. Temporal techniques are incorporated to reflect changes over time, both in the external threat landscape and within the organization. The results are then integrated into a tree structure based on the Factor Analysis of Information Risk (FAIR) methodology. It is a standard quantitative framework for understanding, analyzing, and quantifying information risk in financial terms [8].

C. Visualization and decision support

All previous outcomes are integrated into a web application that offers a complete experience, spanning from data exploration to indicator-driven risk assessment and ultimately to comprehensive cyber risk quantification, ensuring a comprehensive understanding of cybersecurity risks and challenges. This platform consolidates all models and data sources to measure cyber risk for a specific set of entities, for which the project has gathered public data during the evidence collection, or for fictional entities that must be defined using the required model inputs.

Moreover, artificial intelligence (AI) enhances accessibility and usability by providing explanations for diagrams and results. Additionally, the system will incorporate a dedicated recommendation module, which will leverage machine learning models to identify key variables that can be adjusted to achieve the highest return in risk mitigation.

III. Innovations and Key Contributions

DICYME has originated innovative contributions in the three steps of the workflow: data, indicators and cyber risk quantification, and visualization and decision support, developing multiple related modules.

A. Curated datasets

As a result of the evidence collection, valuable datasets have been created using both public and private data sources.

1) Cyber incidents: Nowadays, multiple public databases collect cyber incidents, but they are often limited to specific geographical areas or industry domains due to the origin of the data sources. There is no single, authoritative repository that consolidates all cyber incidents worldwide, primarily because incident data is highly sensitive, and access to such data is fragmented among different stakeholders, each with varying levels of visibility and regulatory obligations. Furthermore, companies have little incentive to disclose cybersecurity incidents unless required by regulations, and even in such cases, reporting is often minimal.

To address these challenges, DICYME has developed automation scripts that aggregate data from several different databases, both OT-specific sources and general cybersecurity repositories [1]. Aggregating data from multiple sources ensures a more comprehensive and representative dataset, mitigating biases that may arise from reliance on a single data provider. The consolidation process uses Natural Language Processing (NLP) techniques to merge duplicate cases, identify entities, and structure the information for efficient analysis.

2) Victim profile: Based on the cyber incidents dataset, impacted entities have been selected, ensuring that the target of each incident corresponds to a single distinguishable entity. The purpose was to build a dataset defining entities through their fundamental characteristics, such as location, size, financial attributes, and reputation. This dataset serves as the foundation for developing models that assess the likelihood of an entity being targeted by threat actors.

Each entity is characterized by variables including headquarters location, industry sector, financial performance indicators (revenue, earnings, and profitability), and organizational attributes such as whether it is publicly traded and its number of employees. While firmographic attributes remain relatively static, the dataset also includes dynamic variables: online reputation, computed through a time-based formula analyzing social media mentions, and victimization metrics, capturing device-related incidents and critical information breaches over time. To ensure robust comparisons, the dataset includes similar entities from the same industry sector that have not reported cyber incidents. The dataset is constructed by integrating multiple free, such as CompaniesMarketCap [15] and Shodan [16], and subscription-based data sources, the latter of which are funded by DeNexus.

3) Intrusion detection systems (IDSs): In addition to external sources, DICYME also proposes indicators based on internal security monitoring systems telemetry, aggregating anonymized IDS data from them. Multiple providers are utilized, with most indicators being applicable across all of them, including Nozomi OT Security Platform [17], Forescout Technologies [18] or Claroty [19]. This dataset captures key security metrics, including asset discovery, vulnerability assessments, and remediation timelines, structured by site, network, Purdue Model levels, severity (common vulnerability scoring system (CVSS)) or cyber asset. Given the sensitivity and confidentiality of such data, the project emphasizes the use of anonymized information wherever possible, ensuring

[1]This includes TI Safe Incident Hub [9], ICS STRIVE [10], KonBriefing [11], Cyber Events Database of the CISSM [12], Hackmageddon [13] & EuRepoC [14]

privacy while preserving its analytical value.

B. Indicators

Using the datasets developed in Section III-A, along with additional datasets utilized in their original published structure, three indicators have been constructed.

1) Attractiveness: A novel cybersecurity metric has been developed to quantitatively assess an entity's likelihood of being targeted by cyber threats. This metric builds upon the *Victim Profile* dataset (see Section III-A2), integrating structured firmographic data with dynamic risk indicators to generate a probabilistic *Attractiveness* indicator. The methodology follows a multi-stage analysis pipeline, built upon three key dimensions:

1) Basal Attractiveness: Represents the inherent characteristics that influence an entity's exposure. Using rule mining (FP-Growth), patterns associated with past cyber incidents are identified within firmographic attributes. A decision tree model further refines this analysis, assigning a probability score in the range $[0, 1]$ to quantify the baseline risk.
2) Online Reputation: Measures the entity's visibility in social media and social networks using a sentiment-weighted engagement formula that evaluates activity and its potential to attract adversarial interest. The resulting score ranges from $[-0.5, 1]$, where positive values indicate strong and favorable engagement, while negative values suggest reputational risks that may increase the likelihood of being targeted.
3) Potential Victimization: Assesses the entity's visibility in underground cybercriminal ecosystems using a buffer that classifies risk into three levels: 0 for entities with neither visible internet-facing devices nor presence in ransomware leaks, 1 for those with either visible devices or leak presence, and 2 for entities with both, indicating significant exposure to cyber threats.

Once these three attributes are computed, they are integrated using a logistic regression model, which balances their contributions into a final *Attractiveness* indicator within the normalized range $[0, 1]$. Empirical validation of this methodology was conducted on a part of the Victim Profile dataset, confirming that higher *Attractiveness* scores correlate significantly with observed cyber incidents. This probabilistic score provides a dynamic measure of an entity's susceptibility to cyber incidents, enabling proactive risk assessment and strategic cybersecurity decision-making.

2) THRACT: A novel indicator has been designed to classify and analyze cyber threat actors over time. This indicator allows organizations to build a customized cybercrime landscape based on publicly available data, such as the encyclopedia of threat actors developed by the Electronic Transactions Development Agency (ETDA) [20]. By aggregating information on actors' identities, locations, motivations, historical activity, targeted sectors, and tools used, DICYME offers a structured framework for assessing cyber threats. The index is composed of three core scores: the Activity Score, which reflects an actor's recent activity; the Capacity Score, which measures their technical capabilities; and the Target Score, which evaluates the level of threat an actor poses to specific organizations based on industry and geographic focus.

This approach offers several key advantages. The index is formulated in a clear and accessible manner, making it usable by both cybersecurity experts and non-specialists. It is dynamic and continuously updated to reflect emerging threats and evolving cybercriminal strategies. Moreover, it is flexible, allowing organizations to incorporate data from multiple sources to refine their threat assessments. By tailoring the analysis to an organization's specific industry and location, the index provides actionable insights, helping businesses enhance their cybersecurity posture and proactively mitigate risks.

3) CVE2TTs: A robust approach to vulnerability assessment involves not only identifying software vulnerabilities (i.e., Common Vulnerabilities and Exposures, CVES) but also understanding their role in the attack lifecycle. The MITRE ATT&CK framework [21] provides a structured taxonomy that classifies adversary tactics and techniques, offering a standardized way to map attacker behaviors. While Common Vulnerabilities and Exposures (CVES) are cataloged in MITRE's database, organizations struggle to prioritize them based on their actual exploitation potential and impact. Attackers leverage vulnerabilities to gain unauthorized access, install malware, escalate privileges, or exfiltrate sensitive data, making it critical to bridge the gap between raw vulnerability data and real-world adversarial behavior.

To address this challenge, DICYME proposes a Two-Layer Neural Network model that automatically maps CVES to MITRE ATT&CK techniques. The model leverages both publicly available vulnerability descriptions and expert knowledge to create labeled training data, particularly enhancing the coverage of the industrial control system (ICS) matrix. By continuously retrieving and processing CVE data, the system updates its mappings dynamically, ensuring that security teams receive up-to-date intelligence on how vulnerabilities align with adversary techniques. This deep learning-driven approach enables organizations to prioritize vulnerabilities more effectively, improving cyber risk management by focusing on those most likely to be exploited in real-world attack scenarios.

C. Cyber risk quantification (CRQ)

The cyber risk quantification framework in DICYME integrates multiple data-driven components to estimate and model the financial and operational impact of cyber threats, leveraging the datasets and indicators developed throughout the project. Additionally, state-of-the-art data and intelligence from the Denexus Knowledge Center (DKC) [22] serve as baselines and input parameters for the simulations. Following the FAIR taxonomy, cyber risk is quantified as the product of Loss Event Frequency and Loss Magnitude, ensuring a structured and comprehensive approach to risk assessment:

1) Loss Event Frequency represents the expected number of cyber attacks an entity may experience within a given timeframe. It is calculated as the product of the Threat Event Frequency and the Vulnerability. The first one is modeled as a Poisson-distributed process, where the average is computed based on a baseline attack rate (derived from industry-specific external reports), incident trends (estimated from the Cyber Incidents dataset), and the *Attractiveness* indicator. The likelihood of a successful attack is computed using the *THRACT*

indicator, the exposure to MITRE ATT&CK techniques, and predefined security profiles, using DKC data.

2) Loss Magnitude quantifies the financial and operational impact of successful cyber attacks, defined by primary and secondary losses. It is computed using probabilistic simulations, such as Gamma and Beta, combined with cyber insurance parameters such as the cost of an hour of forensic investigation or the cost of life. Entity-specific variables are also considered, such as the Revenue or the cost of an hour of downtime.

D. Visualization and decision support

The data, indicators, and CRQ modules are integrated into DICYME's system (see Figure 1), which is a web application that serves as a centralized platform for cyber risk assessment and quantification. This system consolidates the outcomes of the project within an interactive and user-friendly interface, making them also accessible to third-party users.

The unit of analysis within the platform can be either real-world companies from the *Victim Profile* dataset or fictitious entities defined through user-provided parameters. This flexibility allows to assess the cyber risk for some organizations based on their collected empirical data or to simulate risk scenarios by introducing the required inputs, adapting the analysis to specific security concerns, and adding the dynamism of interacting with the models to see how results change. To enhance decision-making, the platform also generates detailed PDF reports containing personalized data and model outputs, providing structured insights in a reusable format that can be shared across teams or external stakeholders.

A key innovation of the platform is the integration of AI-driven explanations, developed using Large Language Models (LLMs), to enhance interpretability. These explanations dynamically analyze and contextualize visualized risk factors and dataset relationships, making cybersecurity risk assessments more transparent and actionable. As explainability in AI-driven decision-making becomes increasingly critical, this novel approach aligns with the growing demand for interpretable models in cybersecurity, offering users a deeper understanding of the insights generated by the system.

IV. Conclusions

The current work of the DICYME project focuses on refining and completing the simulations within the quantification framework, ensuring a robust and accurate cyber risk estimation. Additionally, efforts are being directed toward implementing recommendation measures based on the outputs of the models and simulations. These recommendations will be designed to consider all input factors while also assessing their feasibility and effectiveness.

There are many ways in which this work can be further extended, with the primary modules intended to be integrated into the DeRisk platform while continuously improving and extending their capabilities with other components of the product. Future enhancements may include the incorporation of additional data sources, a more detailed analysis of security posture as observed from the internet, and multiple possibilities for refinement when internal company data is available.

The developed system represents a significant advancement in the field of cybersecurity risk quantification, demonstrating the value of collaboration between academia and industry. The partnership between Rey Juan Carlos University and DeNexus has fostered a productive knowledge transfer, not only enhancing the technological capabilities of cyber risk modelling but also ensuring that these developments are accessible to a wider audience. The findings of this work are being disseminated through publications submitted to leading conferences and journals, contributing to the broader cybersecurity research community. Additionally, the web application will be made publicly available once a stable and demonstrable version is reached, further promoting accessibility and impact.

Acknowledgements

This work has been funded by the Spanish MICIU under the CPP program in the DICYME project (Ref: CPP2021-009025), partially funded by the XMIDAS project (PID2021-122640OB-I00), and supported by DeNexus Inc.

References

[1] T. M. Chen and S. Abu-Nimeh, "Lessons from stuxnet," *Computer*, vol. 44, no. 4, pp. 91–93, 2011.

[2] J. Roldán, J. Boubeta-Puig, J. L. Martínez, and G. Ortiz, "Integrating complex event processing and machine learning: An intelligent architecture for detecting iot security attacks," *Expert Systems with Applications*, vol. 149, p. 113251, 2020.

[3] D. Bhamare, M. Zolanvari, A. Erbad, R. Jain, K. Khan, and N. Meskin, "Cybersecurity for industrial control systems: A survey," *computers & security*, vol. 89, p. 101677, 2020.

[4] H. Kayan, M. Nunes, O. Rana, P. Burnap, and C. Perera, "Cybersecurity of industrial cyber-physical systems: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–35, 2022.

[5] S. N. Hojjati and N. R. Noudehi, "The use of monte carlo simulation in quantitative risk assessment of it projects," *International Journal of Advanced Networking and Applications*, vol. 7, no. 1, p. 2616, 2015.

[6] S. Raychaudhuri, "Introduction to monte carlo simulation," in *2008 Winter simulation conference*. IEEE, 2008, pp. 91–100.

[7] K. Gai, M. Qiu, and H. Hassan, "Secure cyber incident analytics framework using monte carlo simulations for financial cybersecurity insurance in cloud computing," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 7, p. e3856, 2017.

[8] Center for Internet Security, "FAIR: A Framework for Revolutionizing Your Risk Analysis," Sep. 2023. [Online]. Available: https://www.cisecurity.org/insights/blog/fair-a-framework-for-revolutionizing-your-risk-analysis

[9] TI Safe, "Incident Hub," 2024. [Online]. Available: https://hub.tisafe.com/

[10] Industrial Safety and Security Source and Waterfall Security Solutions, "ICSSTRIVE." [Online]. Available: https://icsstrive.com/

[11] KonBriefing, "Cyberattacks, Hacker attacks, Ransomware attacks," 2025. [Online]. Available: https://konbriefing.com/en-topics/cyberattacks.html

[12] Center for International and Security Studies at Maryland, "Cyber Events Database," 2024. [Online]. Available: https://cissm.umd.edu/cyber-events-database

[13] Paolo Passeri, "HACKMAGEDDON," 2025. [Online]. Available: https://www.hackmageddon.com/

[14] EuRepoC project, "The European Repository of Cyber Incidents," 2025. [Online]. Available: https://eurepoc.eu/

[15] CompaniesMarketCap, "Companies ranked by Market Cap," 2025. [Online]. Available: https://companiesmarketcap.com/

[16] Shodan, "Shodan Search Engine," 2025. [Online]. Available: https://www.shodan.io

[17] Nozomi Networks Inc., "Nozomi OT Security Platform," 2025. [Online]. Available: https://es.nozominetworks.com/

[18] Forescout Technologies, "Forescout." [Online]. Available: https://www.forescout.com/

[19] Claroty, "Claroty," 2025. [Online]. Available: https://claroty.com

[20] Electronic Transactions Development Agency, "Threat Group Cards: A Threat Actor Encyclopedia." [Online]. Available: https://apt.etda.or.th/cgi-bin/aptgroups.cgi

[21] The MITRE Corporation, "MITRE ATT&CK," Nov. 2024. [Online]. Available: https://attack.mitre.org/

[22] Denexus Inc., "DeNexus Knowledge Center," 2025. [Online]. Available: https://www.denexus.io/products/derisk/denexus-knowledge-center

CICERO-DARA Prototipo de informes de seguridad para entidades

Celia Fernández Rosales
Fundación I+D del Software Libre (FIDESOL)
cfernandez@fidesol.org

Carmen María Alba Jiménez
Fundación I+D del Software Libre (FIDESOL)
cmjimenez@fidesol.org

Rubén Morillas
Fundación I+D del Software Libre (FIDESOL)
rmorillas@fidesol.org

Rafael Román Aguilar
Fundación I+D del Software Libre (FIDESOL)
rraguilar@fidesol.org

Laura Vigoya
Fundación I+D del Software Libre (FIDESOL)
lvigoya@fidesol.org

Ana González Muñiz
Fundación I+D del Software Libre (FIDESOL)
agonzalez@fidesol.org

Resumen- **DARA presenta una herramienta cuyo objetivo es la generación automática de informes de riesgos basados en la metodología MAGERIT comprensibles y personalizados como apoyo al conocimiento experto. Ante el incremento de incidentes en ciberseguridad y la acelerada evolución tecnológica, la propuesta se orienta a personalizar y simplificar la evaluación de protocolos de seguridad mediante un cuestionario dinámico y el uso de modelos generativos (GPT-4o y GPT-4o-mini). El estudio aborda los desafíos de la adaptación de cuestionarios estáticos, implementando soluciones técnicas que integran modelos de OpenAI. A través de pruebas realizadas en distintos contextos organizacionales, se analizan tiempos de ejecución, calidad del contenido y estructura de los informes, demostrando la viabilidad y eficiencia del enfoque para mejorar la detección, mitigación y gestión de riesgos en entornos de ciberseguridad a través de informes fácilmente comprensibles por el usuario.**

***Index Terms*- ciberseguridad, análisis de riesgos, MAGERIT 3.0, modelos generativos, inteligencia artificial, automatización de informes, cuestionario dinámico personalizado, detección de amenazas, gestión de riesgos, informe explicable**

Tipo de contribución: *Transferencia*

I. Introducción

En el año 2023 se produjeron más de 83 mil incidentes en ciberseguridad según el Instituto Nacional de Ciberseguridad en España (INCIBE). Es un incremento del 24% respecto a 2022[1]. Esto evidencia la vulnerabilidad de los sistemas de información ante ataques cada vez más sofisticados, en un entorno donde los avances tecnológicos, especialmente en inteligencia artificial (IA), generan oportunidades y nuevos desafíos. Por un lado, la IA potencia la capacidad de detección[2] y respuesta ante amenazas, y por otro, puede emplearse de manera fraudulenta para explotar vulnerabilidades[3]. Además, se estudia la aplicación de técnicas de explicabilidad (XAI)[4], para entender el funcionamiento interno de los sistemas, lo que permite desarrollar estrategias más eficaces para contrarrestar los daños. Ante esta dualidad, resulta crucial que las organizaciones cuenten con mecanismos que les permitan evaluar y mejorar sus protocolos de seguridad.

Ante la evolución tecnológica y limitaciones de recursos que enfrentan las organizaciones para mantenerse al día con estos riesgos, se presenta **DARA** (**D**ocumento de **A**nálisis de **R**iesgo **A**utomático), una herramienta diseñada para la generación de informes de riesgo personalizados, cuyo propósito es ser un nexo entre las innovaciones tecnológicas y las empresas. DARA forma parte del Proyecto CICERO (Contramedidas Inteligentes de Ciberseguridad para la Red del Futuro).

DARA, se presenta un marco contextual basado en la literatura actual, que sitúa el problema en su entorno y contextualiza la situación. Posteriormente, se detalla el proceso de diseño del prototipo, enfatizando los desafíos superados durante su desarrollo y las soluciones técnicas y herramientas empleadas. Como resultado de las pruebas realizadas, se comentan los hallazgos obtenidos, se discuten sus implicaciones y se extraen las conclusiones del estudio.

II. Estado del arte

Los avances tecnológicos han impulsado nuevas metodologías en ciberseguridad, pero también han generado desafíos emergentes. En este contexto, la Inteligencia Artificial (IA) y el aprendizaje automático (ML) destacan por su capacidad de detectar y mitigar amenazas de forma autónoma, optimizando sistemas de detección de intrusos (IDS)[7] y plataformas de protección de endpoints (EPP)[8]. Asimismo, el procesamiento de lenguaje natural (NLP) permite desarrollar soluciones accesibles y comprensibles, facilitando la interpretación de riesgos y la toma de decisiones. A pesar de los beneficios, la IA también ha sido explotada en actividades fraudulentas, generando vulnerabilidades en los sistemas. Para contrarrestarlo, se han implementado estrategias avanzadas de protección y respuesta ante incidentes[5]. La investigación actual se enfoca en optimizar la detección y gestión de amenazas con modelos generativos, capaces de evaluar el estado de ciberseguridad de una empresa y generar informes de riesgos. Estas técnicas combinan ML, análisis predictivo y NLP para construir sistemas más eficientes en un entorno digital complejo[6].

III. METODOLOGÍA

En cualquier entidad proteger la información es una tarea esencial. Esta es la motivación de DARA para crear informes que no solo den pautas sobre la seguridad de una organización

sino que sea en un formato cercano y comprensible. La necesidad de resguardar la información ante posibles violaciones motivó al Consejo Superior de Administración Electrónica (CSAE) a desarrollar la **Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información** (MAGERIT)[9]. En **DARA** se usó MAGERIT, como **norma básica** para la protección de la información.

A. Desarrollo del Cuestionario

Una vez establecidas las bases del análisis de riesgos, se procedió al diseño del cuestionario destinado a conocer el estado actual de la seguridad del entorno a analizar. Para lograr un análisis de riesgos completo, se recopiló información específica mediante un cuestionario orientado a evaluar, según **MAGERIT**, el estado de cada categoría y fase del análisis de riesgos.

B. Desafíos del Cuestionario Estático

El diseño de un cuestionario fijo presenta ciertas limitaciones porque el concepto "información relevante" puede variar entre entidades con actividades disímiles. Entre los problemas identificados destacan:

- **Falta de contextualización**: Un cuestionario genérico puede no captar las particularidades de cada empresa.
- **Ambigüedad de términos**: Los términos empleados pueden ser diferentes según el contexto empresarial.
- **Inflexibilidad para profundizar**: El cuestionario no permite realizar preguntas específicas, dando información superficial.
- **Dificultad de adaptación a diferentes realidades**: Las empresas operan bajo normativas y prácticas internas. Un cuestionario fijo resulta poco versátil e irrelevante.

C. Soluciones y Técnicas de Recolección de Datos

Para superar las limitaciones mencionadas, se consideró el desarrollo de un cuestionario dinámico y personalizado para lo que se implementaron las siguientes soluciones:

- **Contextualización mediante OpenAI:** Uso de la plataforma para generar el cuestionario de forma dinámica. Mediante un prompt adaptado a las características de cada entidad y ajustado a MAGERIT, se incorporó el contexto necesario para la adecuación del cuestionario a las particularidades de cada caso.
- **Facilitación en la obtención de respuestas:** Para mitigar la dificultad para responder a las preguntas, se implementaron generadores de respuestas con ejemplos.
- **Generación del informe de riesgos:** Para elaborar el informe final, se emplearon varios modelos de OpenAI basándose en MAGERIT y en los modelos generativos, permitieron obtener conclusiones estructuradas y coherentes sobre los riesgos identificados en la entidad.

D. Herramientas Utilizadas

El desarrollo del estudio se apoyó en **Jupyter Notebook** para la experimentación y el prototipado. Se utilizó **Bokeh** para construir una interfaz interactiva y amigable. Además, se integró **OpenAI** para la generación automática de cuestionarios contextualizados y la creación de informes de riesgos con modelos generativos. Finalmente, se empleó **WeasyPrint** para transformar estos informes en documentos PDF con una estructura clara.

E. Definición de Criterios de Evaluación

Se definieron **criterios de evaluación** que abarcan la **calidad**, **eficiencia** y **estructura** de los informes generados para analizar la capacidad de personalización del cuestionario, la generación de respuestas y la elaboración del informe final. En cuanto a **calidad**, se valoró la precisión, exhaustividad y coherencia del contenido, alineado con MAGERIT, evaluando la correcta identificación de activos, amenazas, salvaguardas y propuestas de medidas correctivas. La **precisión** se midió por el uso adecuado de términos específicos del contexto, la **exhaustividad** con la inclusión de pautas relevantes y la **coherencia** por la representación fiel de las actividades de la entidad evaluada. La **eficiencia** se analizó en términos de tiempo de ejecución de cada fase, aplicando medidas estadísticas para identificar el impacto de la personalización del prompt en el rendimiento del modelo. Para la **estructura**, se examinó la organización lógica del contenido, la segmentación en secciones claras y la presentación del documento en PDF, prestando atención a las diferencias de visualización entre GPT-4o y GPT-4o-mini. Estos criterios se aplicaron en diversos contextos para validar la capacidad del sistema de adaptarse según la entidad, garantizando una evaluación robusta y completa del prototipo.

La metodología implementada en este estudio combina el modelo **MAGERIT** con herramientas avanzadas de análisis y generación de contenido, superando las limitaciones de los cuestionarios estáticos y adaptándose a las particularidades de cada entidad para garantizar información precisa y relevante. El uso de las tecnologías explicadas permitió **la automatización y presentación de los resultados**, logrando un **análisis de riesgos riguroso.** Este enfoque se ajusta al contexto de cada organización, proporcionando una base sólida para la interpretación y aplicación de los resultados.

IV. RESULTADOS

Para evaluar la capacidad de personalización del cuestionario, la generación de respuestas y el posterior informe, se realizaron pruebas en cuatro tipos de organizaciones: **Organización Genérica, Consultoría, Fábrica de Componentes Electrónicos y AENOR.**

El objetivo fue comprobar si el sistema era capaz de adaptar el contenido de las preguntas según el tipo de organización y, posteriormente, personalizar las respuestas de ejemplo. Se consideró que, en un caso real, sería indiferente utilizar el cuestionario genérico o el cuestionario específico, ya que las respuestas difieren significativamente entre diferentes entidades.

A. Selección y Caracterización de Modelos

De los modelos disponibles en **OpenAI**, se seleccionaron aquellos lanzados a partir del 2023 (denominados *gpt-X*) para garantizar el uso de las versiones más recientes. Los detalles relevantes de cada modelo se resumen en la *Tabla I*.

B. Evaluación del Manejo de Tokens y Eficiencia en el Procesamiento

La generación del informe de riesgos para una organización (sin particularizar) se intentó inicialmente con los modelos

gpt-3.5-turbo y gpt-4, evidenciándose que estos presentaban limitaciones para procesar contextos extensos. Se realizó una investigación comparativa que permitió constatar que:

- **GPT-4o**: Utiliza un sistema de representación del contexto más eficiente, utilizando menos tokens que GPT-4. Esto se traduce en una mayor capacidad para manejar textos extensos.
- **Pruebas Realizadas**: Se ejecutaron 8 pruebas con los modelos gpt-4o-mini y gpt-4o, comprobando el manejo de tokens y el procesamiento del prompt completo.

C. Tiempos de Ejecución

Se midieron los tiempos de ejecución en tres acciones: generación del cuestionario, de respuestas de ejemplo y construcción del informe para cada una de las organizaciones en cada modelo. Los resultados y sus análisis estadísticos se resumen en la *Tabla II* y *Tabla III*. Los análisis muestran que **GPT-4o-mini** es más rápido en todas las acciones evaluadas, con tiempos medios de 13.82s, 26.55s y 37.85s para la generación del cuestionario, la de respuestas y la construcción del informe, respectivamente, en comparación con **GPT-4o**, que mostró tiempos medios de 23.03s, 37.00s y 54.43s.

Al examinar la **variabilidad relativa** mediante el coeficiente de variación, se observó que **GPT-4o-mini tiene una dispersión moderada**, con valores aproximados de 11.5% para la generación del cuestionario, 18.2% para la generación de respuestas y 18.9% para la construcción del informe. En contraste, **GPT-4o** presentó una **variabilidad ligeramente mayor**, especialmente en las tareas de generación de respuestas e informes, con coeficientes de 23.8% y 22.8%, respectivamente, mientras que en la generación del cuestionario el valor fue cercano al 10%.

Aunque el tamaño de muestra (n=4 por condición) limita la aplicación de tests de significancia formales, la diferencia en las medias, sugiere que la **personalización de los prompts** tiene un impacto más pronunciado en **GPT-4o-mini**.

D. Análisis del Contenido y Comparación de Informes Generados

Se analizaron dos baterías de preguntas y los informes generados para una **organización genérica**, utilizando GPT-4o y GPT-4o-mini. Se observaron diferencias en la redacción, organización y nivel de detalle:

- **Contenido de la Batería de Preguntas:**
  - **GPT-4o-mini**: Incorpora una sección final con preguntas orientadas a la toma de decisiones.
  - **GPT-4o**: Centrado en recopilación de información, sin énfasis explícito en la toma de decisiones.
- **Comparación de Informes:**
  Los informes generados se evaluaron en términos de estructura, detalle y enfoque. En las Tablas IV y V, se presenta una comparación resumida.

Para las otras tres organizaciones evaluadas, se replicó el mismo proceso. Los resultados obtenidos corroboraron lo observado inicialmente, confirmando las diferencias identificadas en la redacción, la organización de información y el nivel de detalle.

En pruebas específicas, como en el caso del informe generado para una consultora, se observaron diferencias en el enfoque:

- **Informe con GPT-4o**: Enfocado en la seguridad de una consultoría, enfatizando el cumplimiento del Reglamento General de Protección de Datos (RGPD) y la protección de la información en la nube. Resalta la importancia de auditorías de seguridad, actualización de protocolos criptográficos y el uso de tecnologías emergentes para fortalecer la autenticación y detección de amenazas.
- **Informe con GPT-4o-mini**: Presenta un enfoque más estructurado, con secciones que incluyen el cálculo de riesgos y la evaluación del impacto, recomendando herramientas como el Análisis de Impacto al Negocio (BIA) y mapas de riesgo. Además, enfatiza el cumplimiento normativo, sugiriendo la implementación de un Delegado de Protección de Datos (DPO) y la alineación de políticas de seguridad con el RGPD.

Las respuestas a las preguntas se generaron con un asistente de GPT-4o y GPT-4o-mini, aplicados a cada batería de preguntas según correspondiera. En esta fase de la madurez del prototipo no procedía el uso de entidades reales que pudieran responder a los cuestionarios hasta terminar el desarrollo y comprobar el funcionamiento del prototipo y se optó por utilizar estos modelos para generar respuestas para posteriormente, darles uso dentro del prototipo. En este entorno, las respuestas se presentan como guías orientativas para ayudar a los usuarios en la elaboración de sus propias respuestas y servir como referencia en caso de dudas al completar el cuestionario.

Aunque existen **diferencias en la redacción** entre las respuestas generadas por **GPT-4o y GPT-4o-mini**, ambas ofrecen un **contenido bastante homogéneo** en términos de **información y análisis**. No obstante, la **estructura y el estilo** en que se organiza y presenta la información **varía** entre ambos modelos.

V. CONCLUSIONES

Se han estudiado los modelos generativos disponibles en **OpenAI** y se ha concluido el uso de GPT-4o y GPT-4o-mini para la generación de informes por sus capacidades. Los resultados indican que **GPT-4o-mini presenta tiempos de ejecución significativamente menores** en todas las fases evaluadas en comparación con GPT-4o, lo que sugiere una mayor eficiencia en el procesamiento del contenido. A pesar de algunas diferencias en la redacción, ambos modelos produjeron **información homogénea y de calidad,** cumpliendo con los criterios de precisión, exhaustividad y estructura. El estudio también ha **identificado limitaciones**, como la necesidad de **ampliar el tamaño de la muestra** para aplicar pruebas de significancia robustas y la variabilidad observada en la presentación final del informe. Estas observaciones sugieren que la **personalización de los prompts y la selección del modelo adecuado son factores críticos** para optimizar la implementación práctica de la herramienta.

En conclusión, la investigación demuestra la viabilidad de modelos generativos para crear informes de riesgos personalizados en ciberseguridad. La integración de **OpenAI**, **Jupyter Notebook**, **Bokeh** y **WeasyPrint** optimiza la presentación de resultados, generando información precisa y

contextualizada. Actualmente, se comparan los resultados con motores de **LLaMA**, **Gemini** y **DeepSeek**, observando variaciones. Para su validación, se planean pruebas con usuarios finales y expertos en ciberseguridad mediante sesiones controladas y entrevistas semidirigidas, evaluando tanto la funcionalidad como el conocimiento experto.

VI. AGRADECIMIENTOS

El Proyecto CICERO (Contramedidas Inteligentes de Ciberseguridad para la Red del Futuro) es el único proyecto de ciberseguridad reconocido como Red de Excelencia CERVERA, financiado por el Ministerio de Ciencia e Innovación a través del Centro para el Desarrollo Tecnológico Industrial (CDTI), con cargo a los Presupuestos Generales del Estado 2023 y el Plan de Recuperación, Transformación y Resiliencia. Código de Proyecto: CER-20231019

REFERENCIAS

[1] Instituto Nacional de Ciberseguridad (INCIBE), "Infografía balance de ciberseguridad INCIBE 2023," Aug. 2024. [Online]. Disponible:https://www.incibe.es/sites/default/files/2024-08/Infograf%C3%ADa_balance_de_ciberseguridad_INCIBE_2023_0.pdf

[2] M. A. Khalaf and A. Steiti, "Artificial Intelligence Predictions in Cyber Security: Analysis and Early Detection of Cyber Attacks," 2024, doi:10.58496/BJML/2024/006.

[3] S. Oh and T. Shon, "Cybersecurity Issues in Generative AI," in Proc. 2023 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, Republic of, 2023, pp. 97-100, doi: 10.1109/PlatCon60102.2023.10255179.

[4] O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection," IEEE Access, vol. 12, pp. 23954-23988, 2024, doi: 10.1109/ACCESS.2024.3365140.

[5] S. Muhammad and N. A. Mirjat, "Enhancing Cybersecurity with AI: From Anomaly Detection to Threat Mitigation," Bulletin of Engineering Science and Technology, vol. 1, no. 03, pp. 20–39, 2024.

[6] D. Patil, "Artificial Intelligence In Cybersecurity: Enhancing Threat Detection And Prevention Mechanisms Through Machine Learning And Data Analytics," Dec. 07, 2024. [Online]. Diponible: https://ssrn.com/abstract=5057410, doi: 10.2139/ssrn.5057410..

[7] B. S. Ali, I. Ullah, T. Al Shloul, et al., "ICS-IDS: Application of Big Data Analysis in AI-Based Intrusion Detection Systems to Identify Cyberattacks in ICS Networks," J. Supercomput., vol. 80, pp. 7876–7905, 2024, doi: 10.1007/s11227-023-05764-5.

[8] S. K. Mandru, "Machine Learning and AI in Endpoint Security: Analyzing the use of AI and machine learning algorithms for anomaly detection and threat prediction in endpoint security," Journal of Scientific and Engineering Research, vol. 8, no. 3, pp. 264–270, 2021.

[9] Administración Electrónica, "MAGERIT: Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información," [Online]. Disponible:https://administracionelectronica.gob.es/pae_Home/pae_Documentacion/pae_Metodlog/pae_Magerit.html.

Tabla I: RESUMEN DE MODELOS TIPO GPT-X

Modelo	Fecha	Máx.Tokens	Usos Recomendados	Costo (in/out)
gpt-3.5-turbo	03/ 2023	4,09	Chatbots, asistentes virtuales, generación de txt	$0.002
gpt-4	03/ 2023	8,19	Tareas complejas de comprensión y generación de txt	$0.03/$0.06
gpt-4-32k	03/ 2023	32,76	Procesamiento de txts extensos, análisis detallados	$0.06/ $0.12
gpt-4o	05/ 2024	8,19	Procesamiento y generación de txt, imágenes y audio	$0.0025/$0.01
gpt-4o-mini	07/ 2024	4,09	Aplicaciones que requieren eficiencia y rentabilidad	$0.00015/$0.0006

Tabla II: TIEMPOS DE GENERACIÓN GPT-4O-MINI Y GPT-4O

Acción	Genérica		Consultoría		Fábrica		AENOR	
Modelos GPT	**4o-Mini**	**4o**	**4o-Mini**	**4o**	**4o-Mini**	**4o**	**4o-Mini**	**4o**
Cuestionario	15,83 s	23,38 s	12,02 s	20,02 s	13,31 s	23,10 s	14,08 s	25,61 s
Respuestas	27,47 s	29,58 s	22,82 s	32,91 s	22,87 s	35,86 s	33,02 s	49,60 s
Informe	44,98 s	45,49 s	37,70 s	43,41 s	28,09 s	69,98 s	40,62 s	58,80 s

Tabla III: COMPARATIVA DE INFORMES PARTE I

Aspecto	Primer Informe (GPT-4o)	Segundo Informe (GPT-4o-mini)
Título	No posee título específico	"Informe de Riesgos Basado en el Método **Margerit 3.0"**
Estructura	Secciones: activos, amenazas, salvaguardas, impacto residual y cálculo del riesgo residual.	P**asos diferenciados**: identificación de activos, amenazas, salvaguardas, estimación de impacto y riesgo.
Nivel de detalle	General, sin referencias metodológicas específicas.	Basado en Magerit, con enfoque estructurado y metodológico.

Tabla IV: COMPARATIVA DE INFORMES PARTE II

Aspecto	Primer Informe (GPT-4o)	Segundo Informe (GPT-4o-mini)
Amenazas	Enfocado en riesgos tecnológicos	Considera, además, amenazas naturales e industriales
Gestión de Activos	Menciona infraestructura tecnológica, sin un inventario detallado de activos.	Señala ausencia de inventario de activos y propone su implementación.
Evaluación del Riesgo	Analiza riesgo residual y destaca necesidad de mejora continua.	Utiliza matriz de riesgos que diferencia amenazas según probabilidad e impacto.
Plan de Respuesta	Menciona necesidad de actualizar medidas, sin protocolo específico.	Propone plan de respuesta a incidentes detallado, con medidas específicas.

Sesión Investigación II: IA en ciberseguridad

Optimización de la detección de intrusiones mediante estrategias adaptativas con aprendizaje por refuerzo

David Escudero García
RIASC. Universidad de León
descg@unileon.es

Óscar Mogollón-Gutiérrez
Universidad de Extremadura
Escuela Politécnica de Cáceres
oscarmg@unex.es

José Carlos Sancho Núñez
Universidad de Extremadura
Centro Universitario de Mérida
jcsanchon@unex.es

Noemí DeCastro-García
Departamento de Matemáticas. Universidad de León.
ncasg@unileon.es

***Resumen*—La detección de intrusiones de red es una actividad con cada vez más importancia a causa del incremento anual de ataques. En la literatura, el uso de soluciones de aprendizaje automático es uno de los mecanismos más comunes para mejorar el rendimiento de los sistemas de detección. Uno de los principales problemas de estos enfoques está en el desbalanceo de los datos: el volumen del tráfico malicioso es mucho menor que el del normal, lo que dificulta la creación de un modelo eficaz. El uso de modelos ensamblados, que combinan varios modelos individuales, puede incrementar la robustez frente al desbalanceo de los datos. Para intentar mejorar la eficacia de los modelos ensemble frente al desbalanceo, en este trabajo aplicamos aprendizaje por refuerzo para combinar las predicciones individuales de los modelos del ensemble, con el objetivo de mejorar las predicciones respecto a algoritmos clásicos de votación ponderada.**

***Index Terms*—Aprendizaje por refuerzo, Aprendizaje automático, Detección de intrusiones**

Tipo de contribución: *Investigación original (límite 8 páginas)*

I. Introducción

El rápido desarrollo de las tecnologías de Internet, así como la implementación de diferentes tipos de aplicaciones, ha hecho que la protección de los servicios de red sea cada vez más desafiante. Las medidas de seguridad clásicas, como los cortafuegos y los sistemas de control de acceso, son necesarias para proteger la red de ataques externos. Sin embargo, estas herramientas tienen limitaciones a la hora de detectar ataques internos o casos de intrusión complicados. En consecuencia, los sistemas de detección de intrusiones (IDS) se han convertido en un conjunto esencial de defensas diseñadas para detectar actividades maliciosas que evaden las contramedidas de seguridad más simples como los cortafuegos. Los IDS pueden estar basados en host o en red, cada uno con un enfoque diferente, desde el control de inicios de sesión no autorizados hasta el análisis de paquetes de red. [1].

En particular, el uso de herramientas de aprendizaje automático es uno de los enfoques más utilizados en la literatura para mejorar la eficacia de los IDS [2]. Uno de los principales problemas de la aplicación de aprendizaje automático, no solo al problema de la detección de intrusiones [3], sino a otros como la detección de malware [4], es el desbalanceo de los datos: en general, la proporción de tráfico malicioso será significativamente inferior a la del normal, lo que dificulta que el modelo aprenda a distinguirlos eficazmente. Además, la detección de intrusiones requiere una monitorización constante de la red [3] e implica que los modelos deben mantenerse actualizados. El desbalanceo es particularmente negativo en este escenario [5], ya que puede conducir a una infraponderación de las instancias maliciosas.

Para limitar el impacto del desbalanceo, una de las técnicas propuestas es la aplicación de modelos ensemble [6], en los que varios modelos diferentes se entrenan de forma paralela y cuyas predicciones se combinan, generalmente con un algoritmo de votación ponderada [7] que da más importancia a las predicciones de los mejores modelos.

En este trabajo, analizamos la eficacia de la aplicación de aprendizaje por refuerzo (RL) para combinar las predicciones individuales de cada modelo del ensemble y así mejorar la predicción global. Elegimos el aprendizaje porque ha demostrado ser eficaz para mejorar sistemas IDS [8] y tener capacidad para procesar grandes volúmenes de datos, no sólo desde el punto de vista de la detección [9], sino también para seleccionar medidas defensivas de forma dinámica una vez se ha detectado un ataque [10]. Nuestra hipótesis es que la aplicación de aprendizaje por refuerzo puede proporcionar mejores predicciones que los algoritmos de votación al analizar más en profundidad los patrones en las predicciones los modelos individuales. Por ejemplo, uno de los modelos podría distinguir de forma perfecta entre el tráfico normal y un ataque de denegación de servicio, pero ser incapaz de distinguir entre el resto de ataques. Aunque por sí mismo sería poco útil y un algoritmo de votación le podría asignar una baja ponderación en el ensemble, se debería tener en cuenta si la predicción que da suele ser acertada. Por lo tanto, creemos que la adaptabilidad del RL puede hacer que la de detección de intrusiones sea más robusta e inteligente para mitigar las ciberamenazas en evolución.

El resto del artículo se organiza de la siguiente manera. En la Sección II, presentamos los fundamentos del aprendizaje por refuerzo, incluyendo su formulación teórica y su

aplicación en el contexto de la detección de intrusiones. La Sección III revisa trabajos relacionados que han abordado problemas similares, analizando enfoques previos basados en modelos ensemble y técnicas de aprendizaje por refuerzo. En la Sección IV, detallamos la metodología experimental utilizada, incluyendo la descripción de los conjuntos de datos, los modelos empleados y los procedimientos de evaluación. Posteriormente, en la Sección V, se presentan y analizan los resultados obtenidos, comparando el rendimiento del enfoque propuesto con estrategias tradicionales. Finalmente, en la Sección VI, se exponen las conclusiones del trabajo y se discuten posibles líneas de investigación futura.

II. Fundamentos de aprendizaje por refuerzo

El aprendizaje por refuerzo [11] es una rama del aprendizaje automático cuyo objetivo es obtener agentes que reciben información de un entorno y actúan para maximizar una función de recompensa numérica, que guía al modelo para alcanzar el objetivo deseado, ya sea mover un robot o fijar una política de defensa para mitigar un ataque. El aprendizaje por refuerzo se formaliza comúnmente como un proceso de decisión de Markov (MDP), una tupla $< S, A, T, R >$, donde S denota el conjunto de estados, A el conjunto de acciones, T un mapa que determina la transición de un par estado-acción a un nuevo estado y R la función de recompensa que asigna una recompensa a cada acción realizada en cada estado posible.

Siguiendo el ejemplo de la mitigación de ataques, el estado refleja la información del entorno que recibe el agente; como el número de flujos marcados como sospechosos por el IDS de esa IP; las acciones definen cómo el agente RL puede interactuar con el entorno: por ejemplo permitiendo, bloqueando o analizando más en profundidad la conexión; T refleja el cambio de estado que se produce al tomar una acción en un estado concreto. Para el caso de movimiento de un robot, la posición relativa puede ser parte del estado, por lo que moverse en alguna dirección modificará el estado para la siguiente iteración. No obstante, en otros problemas como es el de mitigar ataques, la transición de estados será, en general, completamente desconocida. Finalmente, la recompensa asigna un valor numérico a la acción tomada en un estado concreto, en función de cómo contribuye a la consecución del objetivo del agente. La recompensa guía el comportamiento del agente, así que es importante fijarla de forma adecuada: en el ejemplo de mitigación si sólo se tiene en cuenta si la acción es correcta (bloquear maliciosas y permitir el resto) el agente podría optar por analizar más en profundidad una gran cantidad de conexiones para evitar errores, pero el análisis conlleva un coste adicional y puede no ser factible. Por lo tanto, la recompensa podría incluir un componente que penalice el coste de la medida seleccionada para evitar estar situación.

Según el planteamiento del RL, el agente interactúa secuencialmente con el entorno: en cada momento recibe un estado, selecciona una acción que desencadena una transición a un nuevo estado, recibe un recompensa y repite el proceso hasta que se alcanza el objetivo o se supera un número de iteraciones prefijado. El objetivo del agente es determinar una política, que asigna a cada estado una acción que maximiza la recompensa acumulada. Por lo tanto, la recompensa debe tener una cierta correspondencia con el objetivo que se desea alcanzar. La recompensa acumulada se modela utilizando una función de valor como la función Q de Bellman (Ecuación 1). Esta función es la recompensa acumulada esperada por tomar la acción a en un estado s. π denota la política, R_t la recompensa inmediata por tomar la acción a en el estado s y γ es un factor de descuento: los valores mayores dan más importancia a las recompensas a largo plazo que a las inmediatas.

$$Q(s,a) = E_\pi[R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) | s_t = s, a_t = a] \quad (1)$$

Maximizando la función Q óptima $Q_*(s,a)$, se puede obtener la política óptima $\pi_*(a|s)$. Sin embargo, en muchos casos, T y R no se conocen de antemano [12], por lo que la función Q de Bellman no se puede utilizar directamente y es necesario utilizar algún enfoque, generalmente iterativo, para aproximar la función de valor y obtener la política óptima. En este trabajo hemos optado por utilizar redes Q profundas, que se describen a continuación.

II-A. Redes Q Profundas

Las redes Q profundas (DQN) [13] aplican redes neuronales como aproximadores al aprendizaje Q, que aproxima la función de valor Q de forma iterativa utilizando la ecuación 2. α es un parámetro que controla el tamaño de la actualización. El algoritmo actualiza iterativamente los valores estimados utilizando la estimación del valor de la acción optima en el siguiente estado.

$$\begin{aligned} Q(s_t, a_t) &= (1-\alpha) Q(s_t, a_t) \\ &\quad + \alpha[R_{t+1} + \gamma max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \end{aligned} \quad (2)$$

Para equilibrar la explotación (la selección de las acciones que maximizan Q) y la exploración (selección de acciones nuevas para determinar su eficacia), el aprendizaje Q adopta una estrategia ϵ-greedy durante el entrenamiento. Esta estrategia establece que a_t sea una acción aleatoria con probabilidad ϵ y la acción óptima (según el conocimiento actual) con probabilidad $1 - \epsilon$. El valor de ϵ suele cambiarse dinámicamente durante el entrenamiento, por ejemplo, utilizando un valor más alto al iniciar el entrenamiento y disminuyéndolo a medida que avanza, para hacer uso del conocimiento recopilado para optimizar Q.

El algoritmo se puede implementar como una tabla que registra el valor Q estimado para cada par acción-estado, que se actualiza a medida que avanza el algoritmo. No obstante, si el espacio de estados tiene alta dimensionalidad, la implementación tabular de Q puede ser inviable. En este caso, es necesario utilizar algún tipo de aproximador de Q. Un enfoque muy utilizado es el aprendizaje por refuerzo profundo (DRL) [11], que utiliza una red neuronal como aproximador de la función de valor y es capaz de lidiar con espacios de estados de alta dimensionalidad. En este trabajo utilizamos las DQN [13], ya que han sido utilizadas en otros trabajos centrados en la detección de intrusiones con buenos resultados [12], [14].

En DQN, la red neuronal calcula una estimación de Q para cada estado y selecciona una acción a según la estrategia ϵ-greedy. El valor estimado de Q para la acción se calcula según

la ecuación 3.

$$y = R_t + \gamma max_a Q(s_{t+1}, a) \tag{3}$$

Los parámetros de la red se actualizan mediante descenso por gradiente sobre la función de pérdida cuadrática media definida en la ecuación 4.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - Q(s_i, a_i))^2 \tag{4}$$

Utilizamos la implementación de DQN en el paquete de Python stable-baselines3 [15].

III. Trabajo relacionado

Uno de los principales problemas de la detección de ataques de red es el desbalanceo de los datos [16]: los ataques tienden a representar una proporción relativamente pequeña del tráfico total, y algunos de estos ataques pueden ser raros. Esto significa que será difícil para el modelo identificar estos ataques dada su escasez. La aplicación de modelos ensemble es uno de los principales enfoques para tratar el desbalanceo de los datos [6] debido a su robustez general ante instancias difíciles. Trabajos como [7] aplican ensembles al problema de la detección de intrusiones con buenos resultados: en este trabajo se consigue reducir en un 50 % los falsos positivos utilizando selección bayesiana para eliminar los modelos con peor rendimiento. En [17] se aplica un algoritmo genético junto al ensemble para combinar las predicciones, con una mejora de 0.05 en la F1-score. Esto nos lleva a pensar que el RL también puede mejorar significativamente el rendimiento de un ensemble combinando sus predicciones, dada su mayor capacidad de aprendizaje.

Por otro lado, existen algunos trabajos que aplican RL a problemas de ciberseguridad, en particular para lidiar con ataques de red. La mayoría intenta explotar las capacidades de RL para adaptarse a cambios en la distribución de los datos y aprender con una cantidad limitada de datos etiquetados [9]. Varios trabajos han obtenido buenos resultados al aplicar RL [18], [12] a estos problemas. En [12], se aplican varios modelos RL como clasificadores para detectar intrusiones utilizando los conjuntos de datos NSL-KDD [19] y AWID [20]. En ambos conjuntos de datos, el modelo RL DDQN [21] supera a una serie de modelos ML tradicionales de uso común, como Random Forest o Gradient Boosting, con la ventaja adicional de tiempos de predicción más bajos. En [22], se aplica RL para predecir el volumen de diferentes tipos de tráfico de red (capa de transporte y aplicación) en redes IoT con el fin de detectar ataques de denegación de servicio (DoS). El modelo RL se compara con otros tres algoritmos de predicción de tráfico. RL mejora la F1-score en 0.1 en comparación con el segundo algoritmo con mejor rendimiento, mantiniendo una tasa de falsos positivos limitada.

Otros trabajos utilizan la adaptabilidad de RL e introducen técnicas de preprocesamiento de datos adicionales para guiar el aprendizaje y lograr algún objetivo: generalmente, mejorar la detección en conjuntos de datos desbalanceados. El trabajo en [23] aplica RL a la detección de anomalías de red en un entorno semisupervisado, con solo una cantidad limitada de instancias etiquetadas. El muestreo de instancias en cada episodio se modifica para maximizar el aprovechamiento de las instancias anómalas etiquetadas y la función de recompensa elegida penaliza los falsos positivos. El algoritmo mejora de el área bajo la curva (AUC) hasta en un 10 % en comparación con otros algoritmos de detección de anomalías. En [24], el RL se utiliza para generar instancias sintéticas para entrenar un detector, con el fin de fortalecer los modelos contra ataques de evasión de modelos de aprendizaje automático. Este enfoque mejora la F1-score hasta en un 30 % en presencia de ataques de evasión.

En relación con nuestro enfoque, algunos trabajos en otros ámbitos aplican RL como herramienta para seleccionar los modelos que deben conservarse dentro del ensemble. El ensemble puede estar integrado directamente en el clasificador RL como en [25]. En este caso, los algoritmos utilizan múltiples aproximaciones de DQN (con pesos compartidos en las capas más internas) para producir un conjunto diverso de estimaciones que mejoran la generalización y minimizan la cantidad de datos necesarios para lograr un rendimiento aceptable. Por otro lado, el RL puede aplicarse para mejorar un conjunto de modelos tradicionales. El trabajo en [14] utiliza ensembles para entrenar con datos desbalanceados y aplica RL para seleccionar qué modelos eliminar del ensemble para mantener sólo los modelos más informativos. Esta estrategia supera a otros ensembles especializados en diferentes conjuntos de datos. En este trabajo buscamos no tanto seleccionar modelos como combinar más eficazmente las predicciones de los modelos existentes, asumiendo que si el modelo no aporta información útil, el algoritmo de RL no lo tendrá en cuenta.

IV. Metodología experimental

En esta sección describimos el desarrollo de los experimentos y el análisis de los resultados.

IV-A. Protocolo experimental

Nuestro objetivo es determinar si los algoritmos RL pueden mejorar el rendimiento de los modelos ensemble fusionando las predicciones de cada modelo individual eficazmente. Por lo tanto, el modelo RL se utilizará como clasificador. En este contexto, los parámetros clásicos del MDP se pueden modelar de la siguiente forma:

- Estados S_i: Instancias del conjunto de datos. El modelo RL recibirá como entrada las predicciones apiladas de cada modelo del ensemble. Por lo tanto, para el modelo RL, cada instancia se caracterizará por las probabilidades de pertenecer a cada clase dadas por cada modelo, en lugar de las características originales. Esto se ilustra en la Figura 1. La etiqueta de cada instancia permanecerá inalterada.
- Acciones a_i: las posibles etiquetas a asignar a la instancia. El modelo RL funcionará, por tanto, como un clasificador.
- Recompensa R_i: seguimos la estrategia de trabajos previos [12] y hacemos uso de una recompensa binaria. Su valor será 1 si la etiqueta seleccionada para la instancia es correcta y 0 en caso contrario. Aunque algunos trabajos como [26] hacen uso de recompensas que tienen el cuenta el desbalanceo de los datos, pero en [12] la

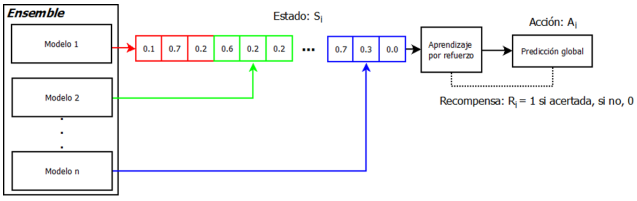


Figura 1. Entrada al modelo RL

recompensa binaria alcanza el mejor rendimiento en el mismo problema que tratamos en este trabajo.

El procedimiento que se sigue en los experimentos para cada conjunto de datos es el siguiente:

- Se entrena el ensemble base y se extraen las predicciones de cada componente del conjunto de los subconjuntos de entrenamiento y prueba
- Las predicciones de cada modelo del ensemble se apilan y se utilizarán como entrada al modelo RL, tanto para entrenamiento como para test.
- Se optimizan los hiperparámetros del modelo RL y se selecciona la mejor configuración.
- Se entrena el modelo RL y se extraen sus predicciones en el conjunto de test.
- Calculamos diferentes métricas para evaluar la calidad de las predicciones tanto del modelo base como del modelo RL.

Para entrenar y probar completamente las configuraciones seleccionadas, seguimos la convención [12] para RL: las instancias se pasan al modelo en lotes aleatorios. Continuamos el entrenamiento hasta que el modelo haya procesado una cantidad de instancias igual al doble de la cantidad de instancias en el conjunto de entrenamiento.

El proceso descrito en esta sección se repite 5 veces con diferentes inicializaciones para conjunto de datos, con el objetivo de determinar si los resultados pueden ser estadísticamente significativos.

IV-B. Modelos e hiperparámetros

Como ensemble base, utilizaremos 4 modelos diferentes: Regresión logística (LR), Árbol de decisión (DT), K vecinos más cercanos (KNN) y perceptrón multicapa (MLP). Utilizamos su implementación en la librería scikit-learn [27] y los combinamos utilizando la estrategia VotingClassifier de la misma librería. En la Tabla I mostramos la configuración de hiperparámetros fijada para cada modelo.

Tabla I
HIPERPARÁMETROS DE LOS MODELOS DEL ENSEMBLE BASE

Modelo	Parámetros	valores
LR	*penalty*	l2
	C	1.0
	max_iter	4000
DT	*criterion*	"gini"
	max_depth	None
	max_features	None
KNN	*n_neighbors*	5
	metric	"minkowski"
	p	2
MLP	*hidden_layer_sizes*	(32, 32)
	activation	"relu"
	max_iter	1000

Para el modelo RL, aplicamos optimización de hiperparámetros siguiendo las recomendaciones de los autores de la librearía [15]. Usamos la biblioteca Optuna [28] para llevar a cabo la optimización. Usamos el muestreador Tree Parzen Estimator [29] y configuramos el número de iteraciones para la optimización en 1000.

La tabla II muestra el espacio de búsqueda de hiperparámetros para el modelo de aprendizaje por refuerzo.

Tabla II
HIPERPARÁMETROS DEL MODELO DQN

Parámetros	valores
learning_rate	[0,00001, 1]
batch_size	[16, 32, 64, 100, 128, 256, 512]
buffer_size	[10000, 100000, 1000000]
exploration_final_eps	[0, 0,2]
exploration_fraction	[0, 0,5]
target_update_interval	[1, 1000, 5000, 10000, 15000, 20000]
learning_starts	[0, 1000, 5000, 10000, 20000]
train_freq	[1, 4, 8, 16, 128, 256, 1000]
subsample_steps	[1, 2, 4, 8]
net_arch	{[64, 64], [256, 256]}

Para ajustar los hiperparámetros, tomamos el 80 % del conjunto de entrenamiento para ajustar cada configuración y el 20 % restante para evaluar su idoneidad.

IV-C. Conjuntos de datos

Utilizaremos dos conjuntos de datos en este trabajo: NSL-KDD [19] y UNSW-NB15 [30]. NSL-KDD contiene características extraídas de tráfico sintético utilizado para evaluación de IDS de DARPA. Esta versión del conjunto datos incluye modificaciones para solventar deficiencias del conjunto de datos KDD99 en el que se basa. El conjunto de datos contiene tráfico normal y 4 tipos de ataque: denegación de servicio, escaneo, acceso remoto y escalada de privilegios. UNSW-NB15 combina tráfico normal real con tráfico malicioso sintético que intenta mejorar la similitud con tráfico real. El conjunto de datos contiene 9 tipos de ataque.

Ambos conjuntos de datos incluyen características que reflejan tanto estadísticas del tráfico (bytes enviados/recibidos) como contenido (número de peticiones HTTP POST y GET).

Para ambos conjuntos de datos, utilizaremos la división en datos de entrenamiento y test proporcionada por los autores. La distribución de clases para ambos conjuntos de datos se muestra en la Tabla III. Como se puede observar, en ambos conjuntos de datos hay un desbalanceo significativo entre las clases, por lo que podemos evaluar eficazmente si el uso de RL permite mejorar el rendimiento de ensembles en situaciones de desbalanceo de los datos.

IV-D. Análisis

Usaremos la F_1-score como métrica de rendimiento de los modelos. Se define en la ecuación 5, donde TP denota verdaderos positivos, FP falsos positivos y FN falsos negativos. Utilizamos esta métrica porque es más informativa que la tasa de acierto al trabajar sobre datos desbalanceados, como es el caso en este trabajo.

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{5}$$

Para aplicar la métrica a problemas multiclase, es necesario obtener un promedio del F_1-score para las diferentes clases.

Tabla III
DISTRIBUCIÓN DE LOS CONJUNTOS DE DATOS

Conjunto de datos	Clase	Frecuencia	
		Entrenamiento	Test
NSL-KDD	Normal	67343 (53.45 %)	9711 (43.07 %)
	DoS	45927 (36.45 %)	7460 (33.08 %)
	Probe	11656 (9.25 %)	2421 (10.73 %)
	R2L	995 (0.79 %)	2885 (12.21 %)
	U2R	52 (0.041 %)	67 (0.88 %)
	Total	125973	22544
UNSW-NB15	Normal	56000 (31.93 %)	37000 (44.93 %)
	Generic	40000 (22.81 %)	18871 (22.92 %)
	Exploit	33393 (19.04 %)	11132 (13.52 %)
	Fuzzers	18184 (10.37 %)	6062 (7.36 %)
	DoS	12264 (6.99 %)	4089 (4.96 %)
	Recon	10491 (5.98 %)	3496 (4.24 %)
	Analysis	2000 (1.14 %)	677 (0.82 %)
	Backdoor	1746 (0.99 %)	583 (0.70 %)
	Shellcode	1133 (0.64 %)	378 (0.45 %)
	Worm	130 (0.07 %)	44 (0.05 %)
	Total	175341	82332

Incluimos los promedios ponderado y macro de F_1-score para dar una visión más completa del rendimiento. El promedio ponderado simplemente suma los conteos de TP, FP y FN para cada clase, por lo que puede dar una valor más optimista si alguna de las clases es muy poco frecuente. El promedio macro computa F_1 de forma independiente para cada clase y calcula la media, lo que conducir a valores más bajos si el F_1-score para alguna de clases minoritarias es bajo.

Además, realizaremos un análisis estadístico de los valores de F_1-score, ponderado y macro, obtenidos para determinar la validez de nuestro enfoque. El proceso se detalla a continuación:

1. En primer lugar aplicamos el test de Kolmogorov-Smirnov de normalidad con la corrección de Lilliefors. Obtenemos resultados no significativos en todos los casos, así que aplicaremos análisis inferenciales parámetricos y utilizaremos la media como medida de centralización.
2. Aplicamos el test t de Student para muestras emparejadas entre los valores de F_1 obtenidos por el ensemble base y las predicciones ajustadas mediante RL. Si hay diferencias significativas, llevaremos a cabo un análisis descriptivo.

Para todos los análisis se fija un umbral de significación $\alpha = 0{,}05$.

V. DISCUSIÓN DE RESULTADOS

En esta sección se presentan y analizan los resultados obtenidos con el enfoque basado en DQN en los conjuntos de datos NSL-KDD y UNSW-NB15. Para ello, se comparan los resultados obtenidos en las 5 repeticiones del experimento y se discute la eficacia de la estrategia de combinación de predicciones en cada caso.

V-A. Resultados en NSL-KDD

En la Tabla IV se muestra el resultado de los análisis estadísticos de los resultados y en la Tabla V se muestran los resultados obtenidos para el conjunto de datos NSL-KDD con distintas inicializaciones de la semilla aleatoria en el entrenamiento del modelo DQN.

El análisis indica que existen diferencias estadísticamente significativas entre los resultados del ensemble base y los del modelo RL tanto para F_1-macro como F_1-ponderado. Por lo tanto, podemos concluir que que las diferencias observadas no son producto del azar, sino que reflejan un impacto real del rendimiento del modelo DQN en la mejora del desempeño.

Se observa que la mejor configuración, correspondiente a la semilla 1062237619, alcanza un valor de F_1-macro de 0.5297 y un F_1-ponderado de 0.7408, superando a las demás inicializaciones. Estos resultados son superiores a los alcanzados por el ensemble base, que obtiene, en el mejor caso, un F_1-macro de 0.4581 y un F_1-ponderado de 0.6817. La mejora media es de 0.06 en el F_1-macro y 0.057 para el F_1-ponderado. Esto refuerza la idea de que una combinación adaptativa de predicciones puede ser más efectiva que la agregación tradicional de modelos en un ensemble, permitiendo a DQN asignar mayor peso a los modelos que realmente aportan información relevante.

Para analizar más en detalle los resultados, en las figuras 2 y 3 se muestran las matrices de confusión de los mejores resultados alcanzados por el ensemble base y el modelo RL respectivamente. Se puede observar que el modelo RL mejora en general al ensemble base, con una mayor tasa de acierto para todas las clases (valores más altos en la diagonal de la matriz). No obstante, la mejora no es de la misma magnitud para todas las clases: DoS y Probe mejoran en 0.06 y 0.07 respectivamente y R2L en 0.13, mientras que U2R (la clase minoritaria) sólo en 0.025. Además, aunque la mejora en R2L es la mayor de todas las clases, sigue teniendo un 87 % de clasificaciones incorrectas. Por lo tanto, aunque hay una mejora significativa en el rendimiento, el modelo no garantiza que la mejora vaya a ser equilibrada o beneficiar particularmente a las clases minoritarias o con peores resultados.

Esta disparidad en la mejora puede deberse a varios factores. Por un lado, el uso de una función de recompensa binaria: no se hace distinción entre predicciones correctas de diferentes clases, por lo que las clases minoritarias tendrán implícitamente menor peso debido a su menor frecuencia. Una función de recompensa que de más peso a las clases minoritarias podría servir para mejorar el rendimiento en clases minoritarias. No obstante en [12] se menciona que esta estrategia no resulta útil con modelos individuales. Por otro lado, para que la predicción global de un ensemble mejore a la de los modelos individuales es necesario que los modelos tengan mayor tasa de acierto que seleccionar una clase aleatoriamente. Como se puede observar en la matriz de confusión (figura 2) la tasa de acierto para las clases R2L y U2R es de 0.00036 y 0.02 respectivamente, por lo que no es posible garantizar que vaya a existir una mejora en estos casos.

En cualquier caso, para las 5 semillas probadas, DQN produce mejoras en todos los casos, para ambos promedios de F_1. No obstante, existe una cierta variabilidad en el rendimiento de DQN, particularmente en F_1-ponderado, donde el modelo llega a 0.7408 en el mejor caso y 0.7177 en el peor, aún evaluando sobre el mismo conjunto de datos. Esta variabilidad observada entre diferentes ejecuciones es un aspecto a considerar en la implementación de este enfoque, ya que pequeñas

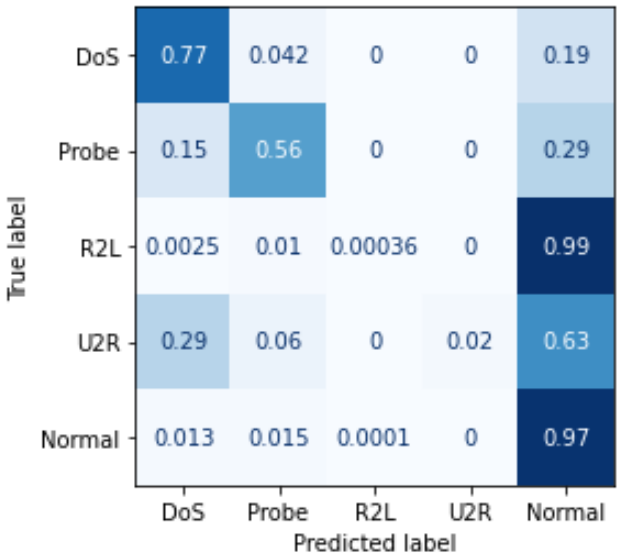


Figura 2. Matriz de confusión norm. del ensemble base en NSL-KDD

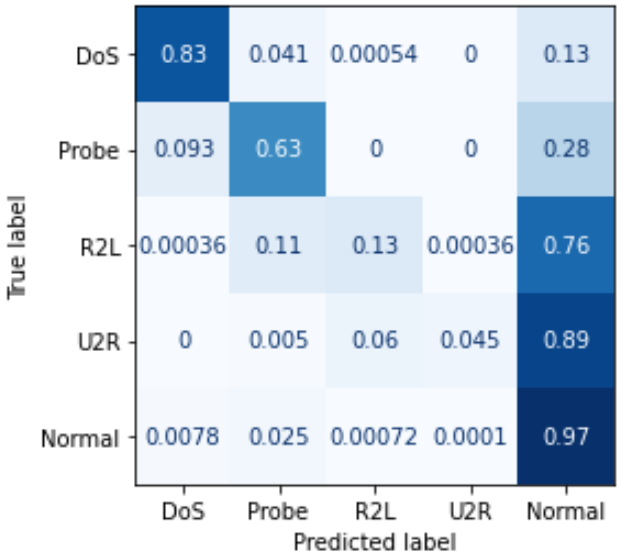


Figura 3. Matriz de confusión norm. del modelo RL en NSL-KDD

diferencias en la política aprendida, ya sea por inicialización o resultados de la optimización de hiperparámetros pueden impactar significativamente el rendimiento del modelo.

Tabla IV
RESULTADOS DEL TEST t DE STUDENT COMPARANDO LAS MÉTRICAS DEL ENSEMBLE BASE Y MODELO RL PARA NSL-KDD

Métrica	Estadístico	ρ-valor
F_1-macro	26.3871	$\rho <$,001
F_1-ponderado	17.3204	$\rho <$,001

V-B. Resultados en UNSW-NB15

En la Tabla VI se muestra el resultado de los análisis estadísticos de los resultados y en la Tabla VII se muestran los resultados obtenidos para el conjunto de datos UNSW-NB15 con distintas inicializaciones de la semilla aleatoria en el entrenamiento del modelo DQN.

Tabla V
RESULTADOS PARA NSL-KDD

Experimento	F_1-macro		F_1-ponderado	
	Base	RL	Base	RL
1062237619	0.4581	0.5297	0.6817	0.7408
2112	0.4508	0.5180	0.6709	0.7379
249651232	0.4467	0.5075	0.6781	0.7372
308118868	0.4512	0.5119	0.6698	0.7177
798844875	0.4444	0.5167	0.6800	0.7322
Media	0.4503	0.5167	0.6761	0.7331

Tabla VI
RESULTADOS DEL TEST t DE STUDENT COMPARANDO LAS MÉTRICAS DEL ENSEMBLE BASE Y MODELO RL PARA UNSW-NB15

Métrica	Estadístico	ρ-valor
F_1-macro	31.7957	$\rho <$,001
F_1-ponderado	52.7972	$\rho <$,001

Al igual que en NSL-KDD, existe una diferencia estadísticamente significativa en los resultados. Además, la magnitud de los valores de t sugiere que la diferencia en desempeño es aún más pronunciada en UNSW-NB15, lo que podría estar relacionado con la naturaleza de los datos y la capacidad del modelo para aprovechar mejor la información disponible.

En el conjunto de datos UNSW-NB15 (Tabla VII), el mejor desempeño se obtiene con la semilla 308118868, alcanzando un F_1-macro de 0.4973 y un F_1-ponderado de 0.7637; superando al mejor resultado del ensemble base, con F_1-macro de 0.4363 y F_1-ponderado de 0.7384. La mejora media es de 0.06 en el F_1-macro y 0.026 en el F_1-ponderado. Al igual que en NSL-KDD, el modelo de RL mejora al ensemble base para todas las semillas evaluadas.

En este caso, la mejora en el F_1-ponderado es menor que en NSL-KDD (0.057). Esto se debe a la diferente distribución de clases en UNSW-NB15 (veáse la Tabla III): este conjunto de datos tiene 10 clases en lugar de 4 como NSL-KDD, con 4 de esas clases con una presencia en torno a 1 % del total. Esto conlleva que un modelo que cometa más fallos en esas clases puede alcanzar un F_1-ponderado relativamente competitivo debido a su escasa frecuencia, pero el valor de F_1-macro obtenido será menor. Esto se puede observar comparando con los valores obtenidos en NSL-KDD: el ensemble base obtiene un F_1-macro de 0.4503, mayor que 0.4302 en UNSW-NB15, pero el F_1-ponderado es de 0.6761 en NSL-KDD y 0.7366 en UNSW-NB15, que es significativamente mayor.

En las figuras 4 y 5 se muestran las matrices de confusión de los mejores resultados alcanzados por el ensemble base y el modelo RL respectivamente. Al contrario que en NSL-KDD, para UNSW-NB15 el modelo RL alcanza mejoras sustanciales en varias clases minoritarias: Shellcode y Worm mejoran en 0.31 y 0.5 respectivamente la tasa de acierto. También hay mejoras en otras clases (0.08 en Normal, 0.05 en Recon), pero con un precio: una disminución de tasa de acierto en 0.15 para Ex ploit y 0.13 para Fuzzers, que constituyen aproximadamente un 13 % y 7 % respectivamente del total de los datos. En este caso, al contrario que en NSL-KDD, el rendimiento en algunas clases se ve disminuido. Aunque en global el modelo RL mejora las predicciones, es necesario tener en cuenta que una mejora en general no implica que el

Tabla VII
RESULTADOS PARA UNSW-NB15

Experimento	F_1-macro		F_1-ponderado	
	Base	RL	Base	RL
1062237619	0.4230	0.4948	0.7353	0.7630
2112	0.4337	0.4943	0.7354	0.7631
249651232	0.4300	0.4936	0.7368	0.7634
308118868	0.4363	0.4973	0.7384	0.7637
798844875	0.4278	0.4921	0.7372	0.7630
Media	0.4302	0.4944	0.7366	0.7632

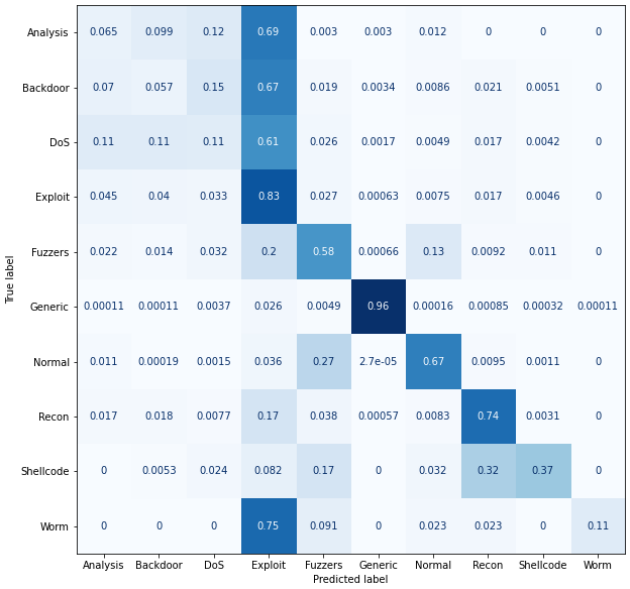


Figura 4. Matriz de confusión norm. del ensemble base en UNSW-NB15

modelo incremente su tasa de acierto sobre todas las clases. Dependiendo de la aplicación, se puede desear una mayor sensibilidad en la detección de ciertos ataques, aún a costa de un rendimiento global algo más bajo.

No obstante la combinación de predicciones del ensemble por RL también mejora significativamente en este conjunto de datos. Esto refuerza la idea de que la aplicación de RL puede ayudar a mejorar el rendimiento predictivo de los ensembles.

VI. CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo, hemos explorado el uso del aprendizaje por refuerzo para la combinación de predicciones en modelos ensemble aplicados a la detección de intrusiones en redes. Nuestra propuesta se fundamenta en la capacidad del RL para adaptar dinámicamente la ponderación de cada modelo dentro del ensemble, superando así las limitaciones de los métodos tradicionales de votación ponderada. Los resultados obtenidos en los conjuntos de datos NSL-KDD y UNSW-NB15 han demostrado que el uso de DQN permite mejorar las métricas de clasificación en entornos con alta desproporción entre clases. En ambos conjuntos de datos, el enfoque basado en RL ha logrado superar a las estrategias de ensemble convencionales, con mejoras significativas en las métrica F_1-macro y F_1-ponderada. Adicionalmente, los experimentos realizados muestran que la estrategia de combinación adaptativa permite un mejor aprovechamiento de los modelos individuales, sin necesidad de una selección previa, lo que refuerza la eficacia

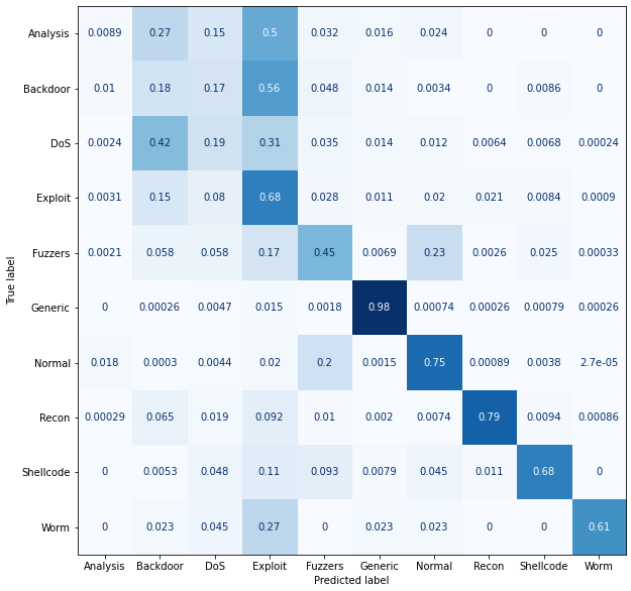


Figura 5. Matriz de confusión norm. del modelo RL en UNSW-NB15

de la propuesta en escenarios de detección de ciberamenazas en evolución.

No obstante, como se ha comentado en la sección V, la mejora general del rendimiento no beneficia necesariamente más a las clases minoritarias o puede incluir una reducción del rendimiento en algunas clases, como ocurre en UNSW-NB15. Por lo tanto, dependiendo del objetivo del ensemble, la aplicación de RL puede requerir un esfuerzo adicional para conseguir que la mejora se ajuste a unos parámetros, ya sea mejor detección en clases minoritarias o ciertas clases en concreto.

Como líneas de trabajo futuras, se explorará la aplicación de otros enfoques de aprendizaje por refuerzo, como los métodos basados en políticas (por ejemplo, PPO o A2C), para evaluar su desempeño en la tarea de combinación de predicciones. En segundo lugar, se investigará el impacto de técnicas de preprocesamiento de datos, como el balanceo de clases mediante generación de muestras sintéticas o la selección de características más representativas, para mejorar aún más la capacidad de generalización del modelo.

AGRADECIMIENTOS

Esta publicación es parte del proyecto *Ciencia de datos para un modelo de inteligencia artificial en ciberseguridad* (C073/23) con clave X67, financiado por la Unión Europea NextGeneration-EU, Plan de Recuperación, Transformación y Resiliencia, a través de INCIBE. También se ha llevado a cabo en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation) - Instituto Nacional de Ciberseguridad (INCIBE) en el Proyecto C108/23 “Detección de Falsificación de Documentos de Identidad mediante Técnicas de Visión por Computador e Inteligencia Artificial”.

REFERENCIAS

[1] K. Sethi, E. Sai Rupesh, R. Kumar, P. Bera, and Y. Venu Madhav, “A context-aware robust intrusion detection system: a reinforcement

learning-based approach," *International Journal of Information Security*, vol. 19, no. 6, p. 657–678, Dec. 2019.

[2] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions," *Artificial Intelligence Review*, vol. 55, no. 1, p. 453–563, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1007/s10462-021-10037-9

[3] A. Shahraki, M. Abbasi, A. Taherkordi, and A. D. Jurcut, "A comparative study on online machine learning techniques for network traffic streams analysis," *Computer Networks*, vol. 207, p. 108836, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2022.108836

[4] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *Journal of Network and Computer Applications*, vol. 153, p. 102526, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.jnca.2019.102526

[5] W. P. Kegelmeyer, K. Chiang, and J. Ingram, "Streaming malware classification in the presence of concept drift and class imbalance," in *2013 12th International Conference on Machine Learning and Applications*. IEEE, Dec. 2013, p. 48–53. [Online]. Available: http://dx.doi.org/10.1109/ICMLA.2013.104

[6] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, p. 221–232, Apr. 2016.

[7] Z. Wu, P. Gao, L. Cui, and J. Chen, "An incremental learning method based on dynamic ensemble rvm for intrusion detection," *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, p. 671–685, Mar. 2022. [Online]. Available: http://dx.doi.org/10.1109/TNSM.2021.3102388

[8] H. Alavizadeh, H. Alavizadeh, and J. Jang-Jaccard, "Deep q-learning based reinforcement learning approach for network intrusion detection," *Computers*, vol. 11, no. 3, p. 41, Mar. 2022.

[9] F. Louati, F. B. Ktata, and I. Amous, "Enhancing intrusion detection systems with reinforcement learning: A comprehensive survey of rl-based approaches and techniques," *SN Computer Science*, vol. 5, no. 6, Jun. 2024.

[10] N. M. Yungaicela-Naula, C. Vargas-Rosales, and J. A. Pérez-Díaz, "Sdn/nfv-based framework for autonomous defense against slow-rate ddos attacks by using reinforcement learning," *Future Generation Computer Systems*, vol. 149, p. 637–649, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.future.2023.08.007

[11] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao, "Deep reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, p. 5064–5078, Apr. 2024. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2022.3207346

[12] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Application of deep reinforcement learning to intrusion detection for supervised problems," *Expert Systems with Applications*, vol. 141, p. 112963, Mar. 2020.

[13] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529–533, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14236

[14] M. Usman and H. Chen, "Emril: Ensemble method based on reinforcement learning for binary classification in imbalanced drifting data streams," *Neurocomputing*, vol. 605, p. 128259, Nov. 2024.

[15] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html

[16] N. Gupta, V. Jindal, and P. Bedi, "Cse-ids: Using cost-sensitive deep learning and ensemble algorithms to handle class imbalance in network-based intrusion detection systems," *Computers & Security*, vol. 112, p. 102499, Jan. 2022.

[17] M. A. Shyaa, Z. Zainol, R. Abdullah, M. Anbar, L. Alzubaidi, and J. Santamaría, "Enhanced intrusion detection with data stream classification and concept drift guided by the incremental learning genetic programming combiner," *Sensors*, vol. 23, no. 7, p. 3736, Apr. 2023.

[18] S. Mohamed and R. Ejbali, "Deep sarsa-based reinforcement learning approach for anomaly network intrusion detection system," *International Journal of Information Security*, vol. 22, no. 1, p. 235–247, Nov. 2022.

[19] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. IEEE, Jul. 2009, p. 1–6.

[20] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, p. 184–208, 2016.

[21] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," 2015.

[22] L. Nie, W. Sun, S. Wang, Z. Ning, J. J. P. C. Rodrigues, Y. Wu, and S. Li, "Intrusion detection in green internet of things: A deep deterministic policy gradient-based algorithm," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 2, p. 778–788, Jun. 2021.

[23] G. Pang, A. van den Hengel, C. Shen, and L. Cao, "Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21. ACM, Aug. 2021.

[24] G. Apruzzese, M. Andreolini, M. Marchetti, A. Venturi, and M. Colajanni, "Deep reinforcement adversarial learning against botnet evasion attacks," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, p. 1975–1987, Dec. 2020.

[25] R. Agarwal, D. Schuurmans, and M. Norouzi, "An optimistic perspective on offline reinforcement learning," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, Jul 2020, pp. 104–114. [Online]. Available: http://proceedings.mlr.press/v119/agarwal20c/agarwal20c.pdf

[26] M. Al-Fawa'reh, J. Abu-Khalaf, P. Szewczyk, and J. J. Kang, "Malbot-drl: Malware botnet detection using deep reinforcement learning in iot networks," *IEEE Internet of Things Journal*, vol. 11, no. 6, p. 9610–9629, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1109/JIOT.2023.3324053

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

[29] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ser. NIPS'11. Red Hook, NY, USA: Curran Associates Inc., 2011, p. 2546–2554.

[30] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*. IEEE, Nov. 2015, p. 1–6. [Online]. Available: http://dx.doi.org/10.1109/MilCIS.2015.7348942

Hacia un Aprendizaje Federado Confiable: Análisis y extensión de PRoT-FL

Idoia Gamiz
Tecnalia, BRTA
Parque Científico y Tecnológico de Bizkaia, 700
48160 Derio (Bizkaia)
idoia.gamiz@tecnalia.com

Cristina Regueiro
Tecnalia, BRTA
Parque Científico y Tecnológico de Bizkaia, 700
48160 Derio (Bizkaia)
cristina.regueiro@tecnalia.com

Eduardo Jacob
Universidad del País Vasco (UPV/EHU)
Dpto. Ingeniería de Comunicaciones
Plaza Ingeniero Torres Quevedo, 1
48013 Bilbao (Bizkaia)
eduardo.jacob@ehu.eus

***Resumen*—El Aprendizaje Federado ha emergido como una solución efectiva para entrenar modelos de inteligencia artificial sin necesidad de compartir datos. Sin embargo, los enfoques tradicionales presentan desafíos en privacidad, robustez, equidad, explicabilidad, responsabilidad y federación, lo que ha llevado al desarrollo del concepto de Aprendizaje Federado Confiable. Este trabajo amplía PRoT-FL, un gestor de entrenamiento federado que optimiza la privacidad y robustez. Para ello, se analiza la confiabilidad de PRoT-FL evaluando los pilares restantes: equidad, explicabilidad, responsabilidad y federación. Además, se propone su extensión con un sistema de recompensas aprovechando la tecnología blockchain en la que se basa. Con ello, PRoT-FL se posiciona como un enfoque de referencia en el ámbito de Aprendizaje Federado Confiable.**

***Index Terms*—Aprendizaje Federado, Confiabilidad, Recompensas, Equidad, Explicabilidad, Responsabilidad, Federación**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Con los continuos avances en la Inteligencia Artificial (IA) [1], las organizaciones buscan cada vez más colaborar en tareas de aprendizaje automático para mejorar sus capacidades predictivas y extraer información valiosa de múltiples fuentes. No obstante, los métodos tradicionales de entrenamiento centralizado requieren compartir grandes volúmenes de datos, lo que plantea serios riesgos para la privacidad. Normativas como el Reglamento General de Protección de Datos [2] desempeñan un papel fundamental en la regulación y aplicación de los requisitos de privacidad. En este contexto, es crucial desarrollar enfoques alternativos que permitan aprovechar el potencial de la IA para entrenar modelos altamente generalizables sin comprometer el cumplimiento de estas normativas.

El Aprendizaje Federado (FL, del inglés Federated Learning) [3] ha surgido en los últimos años como una solución prometedora para el entrenamiento colaborativo de modelos de IA sin necesidad de centralizar los datos de los distintos propietarios, denominados clientes. Sin embargo, este enfoque aún requiere una entidad central de confianza que coordine a los participantes, comúnmente conocida como servidor, S. Sin embargo, la dependencia de un único servidor introduce desafíos en términos de transparencia, equidad y disponibilidad, convirtiéndolo en un punto único de falla.

Además, diversos estudios han señalado que los enfoques tradicionales de FL presentan vulnerabilidades en privacidad y robustez [4]. Por un lado, los ataques a la privacidad pueden producir filtraciones no intencionadas de los datos de entrenamiento a través de la observación de actualizaciones del modelo [5]. Por otro lado, existen ataques adversariales que buscan comprometer la robustez del modelo, dificultando su convergencia mediante manipulación malintencionada [6].

En este contexto, el Aprendizaje Federado Confiable (TFL, del inglés Trustworthy Federated Learning) [7] surge como una alternativa para fortalecer la colaboración en FL, priorizando seis pilares fundamentales: privacidad, robustez, equidad, explicabilidad, responsabilidad y federación. Sin embargo, garantizar estos principios de manera simultánea representa un desafío significativo, ya que la optimización de uno de ellos puede comprometer a los demás.

En este contexto, la solución PRoT-FL [8] se presenta como un gestor de entrenamientos diseñado para garantizar privacidad y robustez simultáneamente, permitiendo la coordinación de múltiples entrenamientos de manera concurrente. PRoT-FL implementa un esquema de FL resistente a ataques de privacidad, asegurando al mismo tiempo la robustez del modelo.

Para lograrlo, el intercambio de modelos se realiza con los modelos cifrados mediante un *Protocolo de Entrenamiento Privado*, que opera a través de canales seguros. Por un lado, este protocolo rompe el vínculo entre un modelo y su generador, lo que lo protege frente a ciertos ataques de inferencia. Además, permite que S descifre y agregue los modelos locales brindando flexibilidad en la selección de reglas de agregación o estrategias para mitigar ataques que comprometan la robustez del modelo. Por otro lado, el protocolo se integra con una red blockchain pública que facilita la auditoría del proceso, garantiza la integridad de los datos y mejora la transparencia. El análisis de PRoT-FL [8] se centró en los pilares de privacidad y robustez. Sin embargo, es necesario extender el análisis a los otros cuatro pilares para gantizar la confiabilidad del sistema.

En este sentido, la principal contribución de este trabajo radica en analizar la confiabilidad del sistema PRoT-FL en base a los seis pilares de TFL. Además, se propone su extensión con un sistema de recompensas basado en blockchain que permite mejorar la cobertura de los pilares no cubiertos.

El resto del artículo se organiza de la siguiente manera. La Sección II explica los fundamentos y antecendentes en los que se basa este trabajo, presentando el protocolo PRoT-FL. La Sección III presenta el análisis de confiabilidad de PRoT-FL en base a sus seis pilares. La Sección IV

ISBN:978-84-10169-61-6

propone un sistema de recompensas para PRoT-FL basado en blockchain. La Sección V compara el nivel de confiabilidad del PRoT-FL con sistema de recompensas con otras soluciones existentes. Finalmente, la Sección VI expone las conclusiones del artículo.

II. Antecedentes

En esta sección, se presentan conceptos necesarios para el desarrollo de la investigación. Se comienza explicando el concepto de TFL y la descripción de sus seis pilares y, posteriormente, se describen los elementos clave de PRoT-FL.

II-A. Aprendizaje Federado Confiable

El concepto de Aprendizaje Federado Confiable (TFL) [7] emerge como una extensión de los principios de la IA Confiable aplicados al ámbito de FL. En la literatura se reconocen seis pilares fundamentales: la privacidad, la robustez, la equidad, la explicabilidad, la responsabilidad y la federación. A continuación, se define cada uno de los pilares. Para equidad, explicabilidad, responsabilidad y federación, se presentan las nociones y las métricas porpuestas por Sánchez et al. [7] para evaluar el cumplimiento de estos pilares.

II-A1. Privacidad: Por privacidad se entiende la protección de los datos de entrenamiento frente a ataques de inferencia y reconstrucción, los cuales buscan extraer información a partir de los modelos compartidos en FL.

II-A2. Robustez: La robustez del modelo hace referencia a su capacidad para resistir ataques que intentan comprometer su integridad, como los ataques de envenenamiento.

II-A3. Equidad: La equidad garantiza un trato justo e igualitario para todos los participantes, promoviendo la justicia en cada etapa del proceso. Se desglosa en diferentes nociones.

- Selección de clientes. Se propone utilizar la "Variación de participación" mediante el cálculo del Coeficiente de Variación (CV) para medir la dispersión en la tasa de participación de los clientes. Con clientes similares, cuanto más dispersa esté la distribución, menos justo será el mecanismo de selección de clientes, y viceversa.
- Rendimiento. Se propone utilizar la "Variación de precisión" para medir la equidad de rendimiento. Se utilizan los modelos locales de los clientes y los datos de prueba de S para medir el *test accuracy*. Cuanto más uniforme sea el *test accuracy* entre los clientes, más justo será el entrenamiento del modelo.
- Nivel de grupo. Se refiere a abordar posibles sesgos demográficos que puedan surgir durante el entrenamiento. Se propone utilizar el "Índice de discriminación" para medir posibles sesgos demográficos.
- Distribución de clases. Se propone que los clientes estimen el desequilibrio de clases de manera local. Para ello, cada cliente calcula el CV de las clases en sus datos antes de iniciar el entrenamiento.

II-A4. Explicabilidad: La explicabilidad se refiere a garantizar procesos transparentes, donde tanto las capacidades como el propósito de los sistemas sean comunicados de manera abierta, y las decisiones sean comprensibles para quienes se ven afectados directa o indirectamente. Se divide en dos nociones.

- Interpretabilidad. Se propone medir tanto el tamaño del modelo como la transparencia del algoritmo, es decir, la complejidad y la interpretabilidad del algoritmo.
- Métodos Post-hoc. Se propone utilizar la simplificación, la importancia de características y la visualización para evaluar la explicabilidad de los modelos.

II-A5. Responsabilidad: Busca garantizar que los sistemas sean verificables y auditables, proporcionando transparencia en cuanto al registro y seguimiento de todas las acciones realizadas. Se desglosa en dos nociones.

- Registro de datos. Se propone completar un documento con toda la información del proceso de FL.
- Monitorización. Se propone realizar una auditoría algorítmica, mediante pruebas funcionales, pruebas de rendimiento, pruebas de aceptación del usuario, o incluso el seguimiento de anomalías o la detección de posibles ataques.

II-A6. Federación: Se refiere a directrices y enfoques arquitectónicos para diseñar sistemas de FL. Se desglosa en diferentes nociones.

- Gestión de clientes. La métrica propuesta es un indicador de la existencia o no de un sistema de gestión de clientes y de un mecanismo para la selección de los mismos.
- Gestión de modelos. Se propone realizar un coversionado de modelos, para alinear las versiones de los modelos locales con sus correspondientes modelos globales agregados.
- Optimización. La elección de métricas dependerá del objetivo principal de la optimización, desde métricas relacionadas con el rendimiento del modelo (si el enfoque está en mejorar la convergencia del modelo) hasta métricas específicas como el uso de ancho de banda, tiempo de entrenamiento por ronda o consumo de energía (si el objetivo es optimizar los recursos).

II-B. PRoT-FL: A privacy-preserving and robust Training Manager for Federated Learning

El sistema PRoT-FL [8], es un gestor de FL que permite orquestar múltiples entrenamientos simultáneamente, garantizando privacidad y robustez, dos de los pilares de TFL. Por un lado, emplea un *Protocolo de Entrenamiento Privado* que permite el intercambio seguro de modelos, evitando su vinculación con los propietarios.

- Privacidad: El hecho de perder el vínculo entre un modelo y su propietario hace que no se puedan aplicar ataques de inferencia y reconstrucción que necesiten información sobre el generador del modelo o modelos consecutivos. El estudio mencionado ya incluye un análisis de privacidad, analizando riesgos y estrategias de mitigación. El mayor riesgo es la recuperación del vínculo. Se concluye que, incluso ante colusión entre S y un máximo de $m-2$ clientes, siendo m el número total de clientes por iteración, la capacidad del adversario para recuperar la identidad del propietario de un modelo depende de las decisiones de los clientes honestos, lo que introduce incertidumbre en el ataque.
- Robustez: PRoT-FL permite que S descifre los modelos antes de la agregación, facilitando la aplicación de reglas

de robustez. El estudio mencionado destaca la adaptabilidad del sistema a diferentes medidas de robustez demostrando que la operativa de PRoT-FL no influye en la precisión del modelo final.

Por otro lado, el *Protocolo de Entrenamiento Privado*, que asegura comunicaciones cifradas fuera de la blockchain, se combina con una red blockchain pública que proporciona trazabilidad sin vincular la identidad de los participantes en ambos entornos. Para ello, se emplean dos contratos inteligentes: `scTrainingManager`, que coordina los entrenamientos, y `scTrainingSession`, que gestiona cada entrenamiento específico. Los entrenamientos se proponen en el primer contrato, que automáticamente genera una instancia del segundo contrato para su gestión. La blockchain registra todo el proceso, almacenando valores *hash* de modelos locales y globales para garantizar su integridad y verificabilidad. Su integración mejora la transparencia y elimina la necesidad de un tercero de confianza, mitigando riesgos asociados a un único punto de fallo.

III. Evaluación de confiabilidad de PRoT-FL

El objetivo de este trabajo es extender el análisis de confiabilidad de PRoT-FL, que ya se había realizado sobre los pilares de privacidad y robustez [8] y que se ha resumido en la Sección II-B, a los pilares restantes: equidad, explicabilidad, responsabilidad y federación. Para evaluar los cuatro pilares restantes, se toman como referencia las métricas definidas en la Sección II-A [7] y se estudia su aplicabilidad.

III-1. Equidad: En cuanto a la selección de clientes, se realiza de forma aleatoria, por lo que todos los clientes tienen la misma probabilidad de ser seleccionados en cada ronda. Es decir, la variabilidad en la tasa de participación entre las rondas debería ser relativamente baja. Respecto al redimiento, S puede medir el *test accuracy* de cada uno de los modelos locales para comprobar si todos los clientes tienen un *test accuracy* similar. Además, se podría extender PRoT-FL para que incluya un sistema de recompensas que permita incentivar a los clientes según el rendimiento de sus contribuciones e, incluso podría utilizarse para la optimización en la selección de clientes, de forma que los clientes que aportan buenas contribuciones tengas más probabilidades de ser seleccionados. En lo que respecta a la equidad a nivel de grupo, actualmente no hay ninguna medida implementada para evitar los sesgos demográficos. Por último, en relación a la distribución de clases, los clientes podrían aportar este dato a S para que lo tenga en cuenta al realizar la agregación. Sin embargo, habría que analizar posibles ataques emergentes por la inclusión de esta información.

III-2. Explicabilidad: La interpretabilidad de un modelo depende del algoritmo concreto y no del sistema en sí mismo. En cuanto a los métodos Post-hoc, en PRoT-FL la blockchain se utiliza como un registro de todo el proceso que visualiza el ciclo de vida completo. Permite la integración de una API que interactúe con la blockchain para obtener consultas y visualizaciones de los datos registrados.

III-3. Responsabilidad: En cuanto al registro de datos, PRoT-FL genera un registro inmutable de todo el proceso de entrenamiento a través de la blockchain, lo que permite automatizar la creación de un documento que refleje la información del entrenamiento a partir de esos registros. Además, en cuanto a la monitorización, PRoT-FL mantiene registros detallados y accesibles, lo que permite realizar auditorías para asegurar que el sistema opera de manera correcta.

III-4. Federación: Para la gestión de clientes, PRoT-FL registra todo el proceso en la blockchain. La selección se hace de manera aleatoria para cumplir con los requisitos del *Protocolo de Entrenamiento Privado* y evitar vincular los modelos con los generadores. En cuanto a la gestión de modelos, S recibe y gestiona todos los modelos locales, lo que le permite implementar un sistema de co-versionado de modelos, fuera de la blockchain. Por último, respecto a la optimización, PRoT-FL ofrece gran flexibilidad para elegir la técnica de optimización más adecuada en cada caso de uso, al no depender de técnicas criptográficas como Computación Segura Multiparte (*SMPC*) o Cifrado Homomórfico (*HE*).

III-5. Resumen: La Tabla I resume el nivel de cumplimiento de PRoT-FL para las nociones clave que sustentan los cuatro pilares adicionales de TFL. El símbolo "✓" indica que PRoT-FL ya hace referencia explícita a ese pilar y, además, propone un mecanismo específico para abordarlo. El símbolo "•" se utiliza cuando el pilar, sea mencionado o no explícitamente en el trabajo, es abordado en la propuesta presentada, al menos parcialmente. El símbolo "○" señala que el pilar no se aborda directamente en los experimentos, pero podría abordarse en otros casos de uso, ya que depende del contexto. Esto es particularmente relevante para la explicabilidad, que está relacionada con el modelo concreto que se entrena. Finalmente, el símbolo "×" denota que el trabajo ni menciona ni ofrece una solución para el pilar. Como se observa en la Tabla I, la equidad es el pilar que requiere de mayores mejoras para conseguir un alto nivel de confiabialidad de PRoT-FL. Como se ha introducido en III-1, un sistema de recompensas podría mejorar la equidad del sistema, al generar recompensas en base al rendimiento y calidad de los resultados de cada cliente, a la vez que abre la puerta a nuevos métodos de selección de clientes optimizados.

IV. Propuesta: Sistema de recompensas para PRoT-FL

En esta sección se presenta el contexto del que surge la implementación de un sistema de recompensas para PRoT-FL. A continuación, se describe el sistema de recompensas y se explica su integración con PRoT-FL. Finalmente, se realiza una validación funcional del sistema.

IV-A. Contexto

Tal como se menciona en la Sección II-B, el uso de blockchain en [8] aporta beneficios significativos en términos de transparencia y que, además, podría contribuir a la equidad.

En particular, se menciona que un sistema equitativo en FL debe asegurar que dispositivos con características similares obtengan los mismos resultados. Asimismo, todos los dispositivos y sistemas deben operar bajo las mismas configuraciones y recibir resultados esperados consistentes. Concretamente, un aspecto fundamental de la equidad es la distribución justa de recompensas entre los participantes honestos. En este sentido, PRoT-FL favorece la equidad al detectar valores atípicos mediante agregación robusta y métricas del modelo, además de utilizar registros intermedios *hash* para permitir a los clientes verificar que todos reciben las mismas versiones

Tabla I
EVALUACIÓN DE CONFIABILIDAD DE PROT-FL: EQUIDAD, EXPLICABILIDAD, RESPONSABILIDAD Y FEDERACIÓN

Nociones		**Descripción**
Equidad		
Selección de clientes	•	Todos los clientes tienen igual probabilidad de ser seleccionados, manteniendo una baja variabilidad en la participación.
Rendimiento	•	S puede medir el *test accuracy* de los modelos locales para verificar su uniformidad entre clientes.
Nivel de grupo	×	Actualmente no hay ninguna medida implementada para evitar los sesgos demográficos.
Distribución de clases	•	Los clientes pueden proporcionar este dato a S para considerarlo en la agregación, pero es necesario analizar posibles ataques emergentes.
Explicabilidad		
Interpretabilidad	○	Depende del entrenamiento concreto.
Métodos Post-hoc	•	La blockchain se utiliza como un registro de todo el proceso que visualiza el ciclo de vida completo. Permite la integración de una API que interactúe con la blockchain para consultar y visualizar los datos registrados.
Responsabilidad		
Registro de datos	✓	Se genera un registro inmutable de todo el proceso de entrenamiento a través de la blockchain, lo que permite automatizar la creación de un documento que refleje la información del entrenamiento a partir de esos registros.
Monitorización	✓	Se mantienen registros detallados y accesibles de las acciones realizadas, permitiendo auditorías para verificar el correcto funcionamiento del sistema.
Federación		
Gestión de clientes	✓	La gestión de clientes se registra en la blockchain, y la selección aleatoria garantiza el cumplimiento del *Protocolo de Entrenamiento Privado*, evitando la vinculación entre modelos y generadores.
Gestión de modelos	✓	S recibe todos los modelos locales, permitiéndole implementar un co-versionado de modelos, fuera de la blockchain.
Optimización	✓	Se ofrece gran flexibilidad para elegir la técnica de optimización más adecuada en cada caso de uso.

del modelo. Además, deja como trabajo futuro la posible inclusión de un mecanismo de incentivos, que haciendo uso de la tecnología blockchain ya empleada, permita fortalecer este principio y motivar contribuciones honestas y efectivas.

IV-B. Descripción

En cada ronda de entrenamiento, S recibe los modelos locales de todos los clientes a través de canales seguros, pero gracias a la definición del *Protocolo de Entrenamiento Privado*, no es capaz de vincular cada modelo con el generador de dicho modelo. Sin embargo, cada cliente registra el valor *hash* de su modelo en la blockchain, lo que permite a S vincular los modelos con la dirección de blockchain asociada al cliente que los generó. Esto posibilita que S recompense a los clientes con un buen modelo sin conocer su identidad fuera de la blockchain, evitando así ciertos ataques de inferencia.

S puede utilizar diferentes métricas para evaluar la calidad de las contribuciones de cada cliente, tales como el volumen de datos utilizados, la precisión del modelo (asumiendo que S posee un conjunto de datos de testeo para verificar la calidad de los modelos recibidos), el consumo energético o el tiempo de computación, entre otras. No obstante, la posibilidad de evaluar de manera fiable estas métricas depende del nivel de confianza o de los mecanismos de verificación externa disponibles, lo cual varía según el tipo de adversario.

Por ejemplo, si cada cliente registra en la blockchain el volumen de datos, S debe asumir que reportan esta información de forma honesta. En ese caso, la equidad del sistema se encuentra condicionada por la integridad de los clientes, lo que podría dar lugar a manipulaciones. No obstante, una posible solución para abordar este problema es el uso de pruebas de conocimiento cero. Estas podrían permitir a los clientes demostrar la veracidad de sus datos sin revelar la información en sí, pero requiere un estudio de su integración.

En cambio, utilizar la precisión de los modelos locales evaluados sobre un conjunto de prueba mantenido por S permite una asignación de recompensas más equitativa y objetiva, al basarse en una medida directa del rendimiento. Sin embargo, este enfoque desplaza el punto de confianza hacia el propio S, ya que los clientes deben confiar en que S emplea un conjunto de prueba imparcial y que no existe colusión entre S y ciertos clientes.

Teniendo esto en cuenta, y una vez definidas las métricas a utilizar, S distribuye recompensas a través de la blockchain, incentivando la participación. Estas recompensas pueden ser en forma de tokens nativos de la red blockchain seleccionada (por ejemplo, ETH en el caso de considerar Ethereum), o, incluso, podría extenderse a la creación de un nuevo token.

Para ello, S establece un valor total que se distribuirá en el entrenamiento y deposita esa cantidad de token en el contrato `scTrainingSession`. Al finalizar el entrenamiento, esta cantidad se reparte entre los clientes de manera proporcional a su evaluación de calidad, lo que significa que la recompensa que recibe cada cliente depende de su desempeño. En otras palabras, la evaluación de calidad de cada cliente se toma en cuenta junto con la cantidad total de recompensa disponible, de modo que aquellos con una mejor evaluación reciban una mayor parte del total, garantizando una distribución equitativa.

IV-C. Integración con PRoT-FL

En esta sección, se especifican las actualizaciones necesarias en la operación de PRoT-FL y en los contratos para integrar este sistema y potenciar la equidad.

IV-C1. Cambios en la operación: La operación completa de PRoT-FL está disponible en [8]. El primer cambio que hay que realizar es en la "Fase 0 - Establecimiento", donde S debe registrar la cantidad total que se va a repartir entre los clientes (la actualización se indica en negrita).

1. S invoca la función publishTraining(). La función genera automáticamente una nueva instancia del `scTrainingSession`. Junto a la propuesta, S especifica cierta información:
   - *input1: dataType.*
   - ...
   - *input6: expectedCapability.*
   - ***input7: totalReward.*** **La cantidad de recompensa que repartirá entre los clientes.**

Despues, todo el proceso se replica con normalidad hasta la "Fase 3 - Validación", en la cual se verifican los modelos locales recibidos a través de los canales seguros. Es en este punto donde se introduce la evaluación de calidad y asignación de recompensas $rewardValue$.

14. S calcula el valor *hash* de los modelos locales cifrados recibidos a través de los canales seguros y verifica si coinciden con los valores *hash* locales registrados en la blockchain. Si todo es correcto, S descifra los modelos locales utilizando su clave privada, sk_S. Si alguno de los valores *hash* no coincide, puede decidir si continuar el entrenamiento sin ese modelo o abortar el protocolo. **Si decide continuar, S evalúa la calidad de los modelos locales recibidos y les asigna un valor $rewardValue$.**

Después, el proceso continúa sin modificaciones con la "Fase 4 - Agregación Robusta", donde se actualiza el modelo global en el paso 15. Es en la "Fase 5 - Etapa de Actualización" donde se requiere un cambio en el paso 16, ya que, para implementar el sistema de recompensas, es necesario registrar la evaluación ($rewardValue$) asignada a cada participante en la blockchain, identificado por el hash de su modelo ($localHash$).

16. Si el entrenamiento no ha finalizado, S calcula el valor *hash* del modelo global actualizado, $h_{\omega^{t+1}}$ e invoca la función *startNextIteration()* para indicar que la siguiente iteración va a comenzar mediante $startNextIteration$. Además, se registra:
    - *input1: globalHash*. El valor *hash* de modelo global calculado previmante.
    - ***input2: localHashes*. Lista de los valores *hash* de los modelos locales evaluados en la iteración.**
    - ***input3: rewardValues*. Cada valor representa la recompensa asignada al modelo cuyo valor *hash* aparece en la misma posición dentro de *localHashes*.**

    Si el entrenamiento ha finalizado, S invoca la función *endTraining()* utilizando su *bcClient*. **Además, registra:**
    - ***input1: localHashes*.**
    - ***input2: rewardValues*.**

IV-C2. Cambios en el `scTrainingManager`: A continuación se muestran las actualizaciones realizadas en el contrato para publicar el entrenamiento, considerando la implementación de referencia de PRoT-FL [8] sobre una red Ethereum y empleando ETH como recompensa.

- Se debe añadir el nuevo parámetro de entrada *totalReward* (cantidad que se va a repartir, en wei) en la función *publishTraining()* así como en el evento *trainingRequest*, para hacer llegar la información a los clientes.

```
function publishTraining(
    DataType memory _dataType,
    AlgorithmInfo memory _algorithmInfo,
    string memory _hashAlgorithm,
    PkeAlgorithm memory _pke,
    string memory _serverAddress,
    uint _expectedCapability,
    uint _totalReward
) public {
    scTrainingSession newTraining = new
        scTrainingSession(_algorithmInfo.
        frac_clients, _algorithmInfo.min_m,
        _totalReward);
    address newTrainingAddress = address(
        newTraining);
    emit trainingRequest(_dataType,
        _algorithmInfo, _hashAlgorithm, _pke,
        _serverAddress,  _expectedCapability,
        newTrainingAddress, _totalReward);
}
```

IV-C3. Cambios en el `scTrainingSession`: Asimismo, se han aplicado las siguientes modificaciones en este contrato para gestionar el entreneamiento.

- Como se puede observar en el código modificado para `scTrainingManager`, para fijar el *totalReward* en el `scTrainingSession` se añade totalReward en el constructor de `scTrainingSession`.

```
constructor(uint _fracClients, uint _minM,
    uint _totalReward) {
    fracClients = _fracClients* 1e4;
    minM = _minM;
    totalReward = _totalReward;
}
```

- Se debe crear un *mapping* del sistema de recompensas (*rewardSystem*) que asocie el valor de evaluación de calidad recibido a cada uno de los clientes. La *Key* es la dirección de blockchain del cliente y el *Value* es el *rewardValue*.

```
mapping (address => uint) public rewardSystem;
```

  Cuando un cliente se registra en el entrenamiento, mediante la función *registerParticipation()*, se inicializa el *rewardValue* del su mapping a 0.

```
rewardSystem[msg.sender]=0;
```

  El *rewardValue* de cada cliente se actualiza con cada iteración finalizada en la que ha participado.
- En la función *startNextIteration()*, ejecutada cuando finaliza una iteración y debe comenzar otra, hay que añadir los nuevos parámetros de entrada *localHashes* y *rewardValues*. Además, esta función debe actualizar los valores de *rewardValues* en el *rewardSystem*. Para ello, se utilizan dos *mappings* ya existentes en la implementación de referencia de PRoT-FL:

```
mapping (uint => address) public participants;
mapping(address => mapping(uint => string))
    public localHashesPerIteration;
```

  El primero guarda la dirección de todos los clientes registrados y el segundo, para cada dirección y cada iteración, el valor *localHash* asociado a dicho cliente. Utilizando esa información, se agregan las siguientes líneas de código en la función *startNextIteration()* para identificar la dirección del cliente asociada a cada *localHash* y, con ella, actualizar el *rewardValue* correspondiente en *rewardSystem*. La actualización consiste en sumar el *rewardValue* en cada iteración para cada cliente.

```
for (uint i = 0; i < localHashes.length; i++) {
    for (uint j = 0; j < n_participants; j++) {
        address participantAddress =
            participants[j];
        if (keccak256(bytes(
            localHashesPerIteration[
            participantAddress][iteration])) ==
            keccak256(bytes(rewardKeys[i]))) {
            rewardSystem[participantAddress] +=
                rewardValues[i];
            break;
        }
    }
}
```

- En la función *endTraining()*, ejecutada cuando finaliza la última iteración y, con ello, la campaña, hay que añadir los nuevos parámetros de entrada *localHashes* y *rewardValues*. Además, esta función debe actualizar los valores de *rewardValues* en el *rewardSystem* por última vez y realizar un reparto proporcional de la recompensa *totalReward* en base al *rewardValue* final de cada cliente.

```
require(address(this).balance >= totalReward, "
    Insufficient balance in contract");
uint totalRewardSystem = 0;
// Calcular el total de las recompensas
for (uint i = 0; i < n_participants; i++) {
    totalRewardSystem += rewardSystem[
        participants[i]];
}
require(totalRewardSystem > 0, "No rewards to
    distribute");
// Repartir las recompensas proporcionalmente
for (uint i = 0; i < n_participants; i++) {
    address participantAddress = participants[i
        ];
    uint participantReward = (rewardSystem[
        participantAddress] * totalReward) /
        totalRewardSystem;
    if (participantReward > 0) {
        (bool sent, ) = payable(
            participantAddress).call{value:
            participantReward}("");
        require(sent, "Failed to send reward");
    }
}
```

- Como el contrato va a gestionar ETH, deben implementarse las funciones básicas. Se define *receive()* para recibir ETH a través de transacciones simples desde una billetera, *deposit()* para enviar ETH de manera explícita y *getBalance()* para consultar el saldo disponible en el contrato en cualquier momento.

```
receive() external payable {}
function deposit() public payable {
    require(msg.value > 0, "Must send ETH");
}
function getBalance() public view returns (uint
    ) {
    return address(this).balance;
}
```

IV-D. Validación funcional

Para validar el sistema de recompensas se ha utilizado Remix IDE [9] y se han desplegado los contratos actualizados de PRoT-FL en la red de prueba Holesky Testnet [10].

Para la validación funcional, se ha fijado una única iteración, con un servidor y tres clientes. En la Tabla II se muestra la información de las cuentas utilizadas tanto para el servidor como para los clientes, así como las direcciones de los contratos actualizados desplegados. Esta información permite verificar en Holesky Etherscan que las transacciones se han realizado correctamente.

Tabla II
DIRECCIONES DE LAS CUENTAS Y DE LOS CONTRATOS DESPLEGADOS EN HOLESKY TESTNET

Cuenta BC: Address
prot-fl-server-0: 0x0b7ce043bB5216Ccc0D3BB3A5A3040Fe2A857E50
prot-fl-client-1: 0x69308C2773550008Bbe06534269751825D646A27
prot-fl-client-2: 0xd508639E8eDc302A75f40570299c33d3DA3d098f
prot-fl-client-3: 0x683741e421Cd0B004adC8B98317DEC49B6940CB5
SC: Address
scTrainingManager:0x6bea5a85ca91b704387556f08a36c171535e936f
scTrainingSession:0xc44e770d2809fa0251720E9cbB72826F60a04b24

Cuando S ha evaluado los modelos locales de cada cliente, asigna los `rewardValues` correspondientes a cada uno. En el ejemplo concreto, S reparte un total de $0,1$ETH y asigna los `rewardValues` de 35, 40 y 25 para los clientes 1, 2 y 3, respectivamente. Los valores que se esperan obtener se muestran en la Tabla III. Además, en la Figura 1 se puede verificar como efectivamente se realiza una transacción de la cantidad correspondiente a cada uno de los clientes.

V. COMPARATIVA CON TRABAJOS RELACIONADOS

Esta sección realiza un análisis comparativo de la confiabilidad de los estudios más relevantes en el ámbito de FL y los compara con PRoT-FL, tras integrar el sitema de recompensas, tomando en consideración el análisis previo de privacidad y robustez [8], así como la evaluación de los demás pilares de confiabilidad que se presenta en la Sección III. La privacidad se evalúa según las técnicas empleadas para resistir ataques de inferencia, clasificadas en *SMPC*, *HE*, Privacidad Diferencial (*DP*), *Desvinculación* u *Otra*. La robustez se analiza según si se usa una técnica específica (*Fija*) o un enfoque flexible (*Flex.*) para proteger el modelo frente a ataques de envenenamiento, lo que posibilita adoptar múltiples enfoques según las necesidades. Además, se considera si los trabajos abordan los pilares de *equidad*, *explicabilidad*, *responsabilidad* y *federación*.

La Tabla IV realiza la comparativa de los diferentes estudios con PRoT-FL extendido. Para ello emplea los mismos símbolos descritos en la Sección III. Además, en el caso de robustez, se utiliza el símbolo "*" para identificar los trabajos que pierden precisión. Por ejemplo, en el caso de trabajos que emplean DP, a pesar de que puedan emplear medidas para preservar la robustez, el uso de DP causa una pérdida de precisión en el modelo [32]. Dado que la robustez implica mantener la precisión, dicha pérdida podría entenderse como que no se logra robustez de manera completa.

En cuanto a los trabajos que abordan privacidad y robustez simultáneamente, todos los trabajos, a excepción de [25], fijan una técnica para hacer frente a los ataques de robustez. Esto los protege frente a ciertos ataques, pero hace que la solución no sea trasladable a ciertos escenarios de ataque. Dependiendo de las propiedades y de la arquitectura de FL, el algoritmo de agregación óptimo varía significativamente. Elegir un método de agregación preciso es crucial para abordar no solo los problemas relacionados con la robustez de los modelos, sino también otros desafíos de FL como la heterogeneidad de los datos, las limitaciones de ancho de banda o las actualizaciones

Tabla III
EVALUACIÓN Y RECOMPENSA PARA CADA UNO DE LOS CLIENTES INVOLUCRADOS

Cliente	**Valor *hash* del modelo**	**rewardValue**	**Recompensa**
prot-fl-client-1	0fedbff077c3d9039518201f9c4914330ea7ba4f8f8942ce7e8d07c265ad06b1	35	0,035 ETH
prot-fl-client-2	71332da02a8fc59daf0d05160baf938eb18a875576d7d199bb5e0c0462929853	40	0,04 ETH
prot-fl-client-3	0d61abfc54da7a1bd77553b2479bce7c4647cbae57f0194cd11970a9dc9f6640	25	0,025 ETH

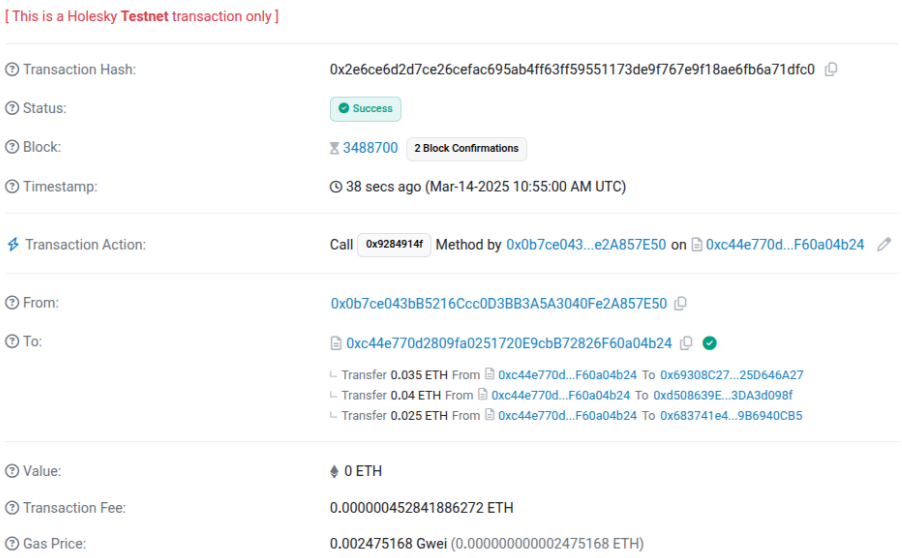


Figura 1. Transacción "0x2e6ce6d2d7ce26cefac695ab4ff63ff59551173de9f767e9f18ae6fb6a71dfc0" donde se realiza la transferencia de ETH a los clientes

Tabla IV
TRABAJO RELACIONADO: APRENDIZAJE FEDERADO CONFIABLE

	Privacidad	**Robustez**		**Equi.**	**Explic.**	**Resp.**	**Fed.**
		Fija	**Flex.**				
[11]	SMPC	✓	×	×	●	×	●
[12]	SMPC	✓	×	●	○	×	×
[13]	HE	✓	×	×	○	×	×
[14]	HE	✓	×	×	●	×	×
[15]	HE	✓	×	×	○	×	●
[16]	HE	✓*	×	×	○	×	●
[17]	DP	✓*	×	×	○	●	●
[18]	DP	✓*	×	×	●	×	×
[19]	DP	✓*	×	×	●	×	×
[20]	DP	✓*	×	×	○	×	×
[21]	Otra	✓*	×	×	○	×	×
[22]	Otra	✓	×	×	○	●	×
[23]	Otra	✓	×	×	○	●	×
[24]	Desvinc.	✓	×	●	○	×	●
[25]	Desvinc.	×	✓	●	○	×	●
[26]	×	✓	×	✓	○	✓	●
[27]	×	✓	×	●	○	✓	●
[28]	SMPC	×	×	×	○	✓	×
[29]	DP	✓*	×	×	○	✓	●
[30]	DP	✓*	×	✓	●	✓	●
[31]	HE	●	×	●	○	✓	×
PRoT-FL: Desvinc.		×	✓	✓	●	✓	✓

asíncronas. Por lo tanto, la flexibilidad sirve tanto para conseguir que el sistema sea resiliente a ataques de envenenamiento, abordando el pilar de robustez, como para lograr agregaciones más óptimas, impulsando el pilar de federación.

Por otro lado, los trabajos que se centran en la técnica de *Desvinculación* tienen el potencial de ofrecer una mayor flexibilidad en la selección de reglas de agregación, ya que no dependen de operar sobre datos cifrados o perturbados. Sin embargo, los trabajos encontrados [24], [25] no explican cómo evitar la dependencia de una entidad central para facilitar la coordinación entre los participantes, lo que constituye un punto de fallo en el sistema y no logra abordar el pilar de responsabilidad y federación.

Para abordar los problemas de equidad, explicabildiad, responsabilidad y federación, destacan principalmente los trabajos basados en blockchain. En cuanto a la equidad, algunos trabajos aportan un valor al dato o un sistema de reputación que podría ser utilizado para ofrecer recompensas a los que contribuyen al proceso y así fomentar la equidad en un futuro [12], [24], [25]. Resaltan los trabajos basados en blockchain, ya que son capaces de ofrecer mecanismos de incentivos para fomentar la equidad [26], [27], [30], [31]. Lo mismo ocurre con la responsabilidad. Aunque algunos trabajos logren aportar cierta responsabilidad mediante la verificación de algunos procesos intermedios [17], [22], [23], para lograr total transparencia y monitorio del proceso y abordar el pilar de responsabilidad sobresalen las soluciones basadas en blockchain [26], [27], [28], [29], [30], [31]. Respecto a federación, aunque algunas soluciones basadas en SMPC o HE optimicen la federación, estas mejoras suelen compararse con otras soluciones criptográficas, por lo que siguen siendo limitadas para una federación a gran escala. Por último, la blockchain facilita la explicabilidad mediante la visualización. Sin embargo, la explicabilidad en FL depende del modelo entrenado más que del sistema en sí, por lo que su estudio es muy relevante para estudiar cada caso de uso específico, pero no para estudiar el sistema de PRoT-FL en su conjunto.

Como se puede observar, PRoT-FL es el único trabajo que ofrece simultáneamente privacidad y robustez de manera integral, al tiempo que permite una selección flexible de las

reglas de agregación y una trazabilidad del proceso para fomentar otros pilares de la confiabilidad como la equidad, la responsabilidad, la explicabilidad y la federación.

VI. Conclusiones

Este trabajo amplía PRoT-FL [8], un gestor de entrenamiento federado previamente validado en términos de privacidad y robustez. Se analizan y validan otros aspectos claves del Aprendizaje Federado Confiable: la equidad, explicabilidad, responsabilidad y federación. Además, se propone, implementa e integra un sistema de recompensas que promueve la equidad, el pilar menos cubierto por la solución original.

La validación funcional confirma la correcta integración del sistema de recompensas en PRoT-FL, mientras que el análisis de confiabilidad posiciona a PRoT-FL como un enfoque de referencia en Aprendizaje Federado Confiable. En comparación con otros trabajos en este campo, PRoT-FL es el primero en ofrecer una solución integral que combina privacidad y robustez con una selección flexible de reglas de agregación, y que refuerza otros pilares clave de la confiabilidad, como la equidad, la responsabilidad, la explicabilidad y la federación.

A partir de los resultados obtenidos en este estudio, se identifican varias líneas de trabajo futuro para PRoT-FL:

- Mejorar otros aspectos identificados en el análisis de confiabilidad de PRoT-FL mediante futuras nuevas extensiones del PRoT-FL actual.
- Validar el sistema de recompensas frente a posibles ataques, así como estudiar las estrategias de mitigación correspondientes. Esto incluye, entre otros aspectos, la manipulación de métricas tanto por parte de los clientes como de S, tal como se menciona en este trabajo.
- Explorar otras alternativas, como permitir que S envíe ETH directamente a los clientes, implementar un token ERC-20 propio que otorgue privilegios dentro del sistema, o explorar mecanismos de reputación en lugar de recompensas monetarias.

Agradecimientos

Este trabajo ha sido parcialmente financiado por TECNALIA y la Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU) mediante el contrato con referencia PIFTEC21/05, y por el Ministerio de Ciencia de España a través del proyecto AIBioSurv-Tech (Biovigilancia mediante Inteligencia Artificial (IA) en la era post-COVID: Implicaciones para la arquitectura y la ciberseguridad) con referencia TED2021-129975B-C22.

Referencias

[1] Padhi, B. K., Nayak, S., y Biswal, B.: “Machine learning for big data processing: A literature review”, en Int. J. Innov. Res. Technol, vol. 5, n. 7, pp. 359-368, 2018.

[2] GDPR: “General data protection regulation”, 2018. URL: https://gdpr-info.eu.

[3] McMahan, B., Moore, E., Ramage, D., et al.: “Communication-efficient learning of deep networks from decentralized data”, en *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273-1282, 2017.

[4] Lyu, L., Yu, H., Ma, X., et al.: “Privacy and Robustness in Federated Learning: Attacks and Defenses”, en *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, n. 7, pp. 8726-8746, 2024.

[5] Melis, L., Song, C., De Cristofaro, E., et al.: “Exploiting Unintended Feature Leakage in Collaborative Learning”, en *2019 IEEE Symposium on Security and Privacy*, pp. 691-706, 2019.

[6] Fang, M., Cao, X., Jia, J., et al.: “Local model poisoning attacks to byzantine-robust federated learning”, en *Proceedings of the 29th USENIX Security Symposium*, pp. 1623-1640, 2020.

[7] Sanchez., P.M., Huertas, A., Xie, N., et al.: “ FederatedTrust: A solution for trustworthy federated learning”, en *Future Generation Computer Systems*, vol. 152, pp. 83-98, 2024.

[8] Gamiz, I., Regueiro, C., Jacob, E., et al.: “PRoT-FL: A privacy-preserving and robust Training Manager for Federated Learning”, en *Information Processing & Management*, vol. 62, n. 1, pp. 103929, 2025.

[9] Bienvenido a la documentación de remix. URL: https://remix-ide.readthedocs.io/

[10] Holesky Testnet. URL: https://holesky.ethpandaops.io/.

[11] Dong, C., Weng, J., Li, M., et al.: “Privacy-preserving and Byzantine-robust federated learning”, en *IEEE Transactions on Dependable and Secure Computing*, pp. 1-16, 2023.

[12] So, J., Guler, B., Avestimehr, A. S.: “Byzantine-resilient secure federated learning”, en *IEEE Journal on Selected Areas in Communications*, vol. 39, pp. 2168-2181, 2021.

[13] Ma, X., Zhou, Y., Wang, L., Miao, M.: “Privacy-preserving Byzantine-robust federated learning”, en *Computer Standards Interfaces*, vol. 80, pp. 103561, 2022.

[14] Ma, Z., Ma, J., Miao, Y., et al.: “ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning”, en *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1639-1654, 2022.

[15] Liu, X., Li, H., Xu, G., et al.: “Privacy-enhanced federated learning against poisoning adversaries”, en *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4574-4588, 2021.

[16] Bai, Y., Fan, M.: “A method to improve the privacy and security for federated learning”, en *IEEE ICCCS*, pp. 704-708, 2021.

[17] Gao, X., Fu, S., Liu, L., Luo, Y.: “BVDFed: Byzantine-resilient and verifiable aggregation for differentially private federated learning”, en *Frontiers of Computer Science*, vol. 18, 2024.

[18] Zhang, Z., Hu, R.: “Byzantine-robust federated learning with variance reduction and differential privacy”, en *2023 IEEE Conference on Communications and Network Security (CNS)*, pp. 1-9, 2023.

[19] Ma, X., Sun, X., Wu, Y., Liu, Z., Chen, X., Dong, C.: “Differentially private Byzantine-robust federated learning”, en *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, pp. 3690-3701, 2022.

[20] Naseri, M., Hayes, J., De Cristofaro, E.: “Local and central differential privacy for robustness and privacy in federated learning”, en *CoRR*, abs/2009.03561, 2020.

[21] Xu, S., Xia, H., Liu, P., et al.: “FLPM: A property modification scheme for data protection in federated learning”, en *Future Generation Computer Systems*, vol. 154, pp. 151-159, 2024.

[22] Tang, X., Shen, M., Li, Q., et al.: “PILE: Robust privacy-preserving federated learning via verifiable perturbations”, en *IEEE Transactions on Dependable and Secure Computing*, pp. 1-18, 2023.

[23] Zhang, Z., Li, J., Yu, S., et al.: “SafeLearning: Secure aggregation in federated learning with backdoor detectability”, en *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 3289-3304, 2023.

[24] Jebreel, N. M., Domingo-Ferrer, J., Blanco-Justicia, A., et al.: “Enhanced security and privacy via fragmented federated learning”, en *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-15, 2022.

[25] Domingo-Ferrer, J., Blanco-Justicia, A., Manjon, J., Sanchez, D.: “Secure and privacy-preserving federated learning via co-utility”, en *IEEE Internet of Things Journal*, vol. 9, pp. 3988-4000, 2022.

[26] Jin, S., Li, Y., Chen, X., et al.: “Blockchain-based fairness-enhanced federated learning scheme against label flipping attack”, en *Journal of Information Security and Applications*, vol. 77, p. 103580, 2023.

[27] Ma, X., Xu, D., Wolter, K.: “Towards blockchain-enabled decentralized and secure federated learning”, en *Information Sciences*, vol. 665, p. 120368, 2024.

[28] Houda, Z. A. E., Senhaji Hafid, A., Khoukhi, L.: “MITFed: A privacy-preserving collaborative network attack mitigation framework based on federated learning using SDN and blockchain”, en *IEEE Transactions on Network Science and Engineering*, pp. 1-17, 2023.

[29] Yan, X., Miao, Y., Li, X., et al.: “Privacy-preserving asynchronous federated learning framework in distributed IoT”, en *IEEE Internet of Things Journal*, p. 1, 2023.

[30] Zhang, Y., Tang, Y., Zhang, Z., et al.: “Blockchain-based practical and privacy-preserving federated learning with verifiable fairness”, en *Mathematics*, vol. 11, 2023.

[31] Weng, J., Weng, J., Zhang, J., et al.: “DeepChain: Auditable and privacy-preserving deep learning with blockchain-based incentive”, en *IEEE Transactions on Dependable and Secure Computing*, pp. 1-1, 2019.

[32] Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A.: “The limits of differential privacy (and its misuse in data release and machine learning): Differential privacy is not a silver bullet for all privacy problems”, en *Communications of the ACM*, vol. 64, pp. 33-35, 2021.

Design of a GenAI-Powered Platform for Cyber Threat Intelligence Analysis

Jaime Fúster, Roberto González, Giuseppe Siracusano, Davide Sanvito, Roberto Bifulco
NEC Laboratories Europe
first.last@neclab.eu

***Abstract*—In this paper, we present the design of a modular and scalable platform designed to automate the collection, processing, and analysis of cyber threat intelligence (CTI). The platform leverages a combination of generative AI (genAI) and graph-based analytics to enable organizations to efficiently process large volumes of heterogeneous threat data. Our system is capable of autonomously collecting intelligence from structured and unstructured sources, enriching the data, and producing actionable insights in real-time. The platform's architecture is organized into distinct layers focused on intelligence collection, data processing and enrichment, and user interaction and decision support. With over one billion STIX objects processed in its knowledge graph, the platform empowers cybersecurity experts and IT professionals to navigate complex threat environments and make informed decisions quickly. Additionally, its modular design ensures seamless integration with legacy systems and facilitates future expansion. We also discuss the need for enhanced data provenance verification, as the growing adoption of automated CTI platforms may lead to adversarial attempts to pollute open-source intelligence. This work represents a step forward in streamlining CTI workflows and improving cyber defense capabilities.**

***Index Terms*—Cyber Threat Intelligence, Cybersecurity**

Type of contribution: Work-in-progress

I. Introduction

In the current global digital landscape, organizations are increasingly interconnected and reliant upon information technologies. This interconnectedness brings numerous operational advantages, such as efficiency, speed, and innovation. However, it also significantly expands the attack surface, making organizations more vulnerable to a wide array of cybersecurity threats [1]. Cyber incidents, ranging from data breaches [2, 3, 4], to complex, targeted cyber-attacks [5], have evolved rapidly in frequency, complexity, and potential impact, posing substantial financial, operational, and reputational risks. As cyber threats continue to escalate, it is imperative for organizations to proactively understand, anticipate, and mitigate potential cyber risks to ensure robust and resilient security postures.

Cyber Threat Intelligence (CTI) has emerged as a critical discipline within cybersecurity [6], designed specifically to address the increasing sophistication of threats and the dynamic nature of the global cyber threat landscape. CTI involves systematically collecting, analyzing, and disseminating information related to cyber threats, attackers, methods, and potential vulnerabilities. By leveraging CTI, organizations can gain deeper insights into the threat landscape, enabling them to make informed, timely decisions that enhance their defensive strategies and response capabilities. Effective CTI practices not only help organizations respond reactively to immediate threats but also proactively anticipate emerging risks, thus strengthening their overall cybersecurity resilience.

One of the key challenges facing modern cybersecurity strategies is the sheer volume and variety of available threat data [7, 8]. Information relevant to cybersecurity originates from multiple sources, including open-source intelligence (OSINT), private threat feeds [9], dark web forums, vulnerability databases, and social media channels. While these diverse sources provide valuable context and data points, the volume and complexity of the data often overwhelm traditional analysis methods. Manual or semi-automated approaches to threat analysis often fall short in terms of speed, accuracy, and scalability, leading to gaps in coverage and potentially critical delays in threat detection and response [10].

Moreover, the global nature of cyber threats complicates the effective use of traditional security solutions. Cyber adversaries frequently leverage sophisticated techniques, exploiting vulnerabilities that cross geographic and organizational boundaries. Attackers utilize tools and strategies that evolve rapidly, necessitating an equally adaptive and dynamic response. Organizations require platforms capable of integrating and analyzing diverse, globally distributed threat data to identify relevant patterns, anticipate potential attacks, and develop effective mitigation strategies quickly and accurately.

Recent advancements in artificial intelligence (AI), particularly in Natural Language Processing (NLP), Large Language Models (LLMs) [11], and graph-based analysis (GraphAI) [12, 13], offer promising solutions to address these challenges. These advanced technologies enable automated, accurate, and scalable processing of vast amounts of structured and unstructured threat data. Utilizing AI-driven techniques, CTI platforms can effectively sift through large datasets to identify meaningful insights, extract critical information such as indicators of compromise (IoCs), and detect subtle trends or patterns indicative of emerging threats.

An integrated and automated CTI platform that harnessed AI-driven analytics would significantly enhance an organization's capability to process and interpret threat intelligence efficiently. Such a platform would enable security analysts to transition from reactive firefighting to proactive threat management, empowering them to act based on comprehensive situational awareness rather than fragmented or incomplete information. By providing actionable intelligence in real-time or near-real-time, these advanced solutions would improve the speed, accuracy, and effectiveness of decision making, thereby substantially strengthening the defensive posture of an organization.

In this paper, we present the design of an advanced, auto-

mated CTI analysis platform specifically designed to address the aforementioned challenges and improve organizational cybersecurity resilience. The proposed platform integrates state-of-the-art technologies including LLMs, graph-based threat analytics, and autonomous agents for data acquisition and enrichment. The comprehensive approach described in this paper ensures not only rapid identification and analysis of threats but also provides intuitive interfaces and actionable outputs that significantly streamline operational decision-making.

The remainder of this paper is organized as follows: Section II reviews related work, Section III outlines the background and key concepts, Section IV describes the design of the proposed platform, Section V discusses the main findings and outlines future work, and Section VI presents the conclusions.

II. Related Work

Recent advancements in the enhancement of CTI workflows with genAI have focused on automating data extraction, structuring intelligence using knowledge graphs (KGs), and improving real-time threat detection capabilities. Large language models (LLMs) such as GPT-4 and Mistral-7B are increasingly integrated into CTI processes to ingest and process unstructured threat data, extract meaningful indicators, and generate actionable intelligence [14, 15, 16]. Automated reporting systems, such as AGIR, combine template-based generation with LLMs to produce clear, accurate summaries while minimizing hallucinations [17]. Knowledge graph-driven approaches, like AttacKG+ and CTINexus, structure CTI reports into queryable representations, enhancing intelligence retrieval and scalability, particularly in resource-constrained environments [16, 18]. In dynamic threat detection, models employing contrastive learning and semi-supervised techniques, such as GAN-BERT, have demonstrated improved identification of threats in open-source intelligence feeds [19]. Furthermore, simulations powered by genAI, using techniques such as GANs and variational autoencoders, have been employed to model attacker behavior and refine defensive strategies proactively [20]. Despite these advancements, key challenges remain, including scalability issues, adversarial risks, and ethical concerns related to the dual-use of genAI for malicious activities [21, 22]. Overall, current research is advancing toward more scalable, interpretable, and adaptive genAI solutions for CTI, addressing reliability, security, and ethical concerns in equal measure.

CTI platforms, have evolved to address the growing complexity of cyber threats by enabling the aggregation, analysis, and sharing of intelligence across organizations. Industry-standard solutions such as MISP [23] have emerged as key open-source platforms, facilitating structured intelligence sharing through standardized formats like STIX/TAXII. Comparative studies evaluate platform performance based on critical criteria such as interoperability, threat detection accuracy, and scalability [24, 25, 26], while researchers are exploring novel architectures that incorporate graph databases for advanced correlation [27, 28], federated learning for privacy-preserving intelligence sharing [29, 30], and blockchain for trust-enhanced threat data exchange [29, 31]. Although automation has advanced significantly [32, 33, 34], challenges persist in standardization, usability, long-tail threat detection, and ensuring robust implementation of the intelligence cycle [25, 35, 36]. These gaps underscore the need for continued empirical studies to enhance the effectiveness of CTI platforms across diverse operational environments.

III. Background

This section provides definitions of key terms and concepts fundamental to understanding Cyber Threat Intelligence (CTI) and the technologies employed to enhance cybersecurity practices:

STIX: Structured Threat Information Expression is a standardized language developed to enable the consistent representation and sharing of CTI. STIX systematically describes IoCs, threat actors, vulnerabilities, incidents, and tactics, techniques, and procedures (TTPs) in a structured, machine-readable format, enhancing interoperability and automated information sharing across cybersecurity tools and platforms.

Cyber Threat Alliance (CTA): A nonprofit organization dedicated to enhancing cybersecurity through collaboration and timely sharing of actionable threat intelligence among global cybersecurity practitioners and stakeholders.

Generative AI (genAI) Agents: Advanced artificial intelligence agents powered by LLMs, such as the GPT family, DeepSeek or Google Gemini, designed to autonomously analyze, interpret, and extract meaningful insights from complex and unstructured textual data.

IV. Platform Design

This section describes our comprehensive approach to designing an advanced, integrated CTI platform. We will begin by outlining our core objectives and requirements, proceed with a general overview of the platform architecture and schema, briefly describe the different platform components, and finally provide an in-depth example of a specific, representative component, demonstrating how it contributes practically to the platform's overall effectiveness and value.

A. Platform Objectives and Requirements

In developing our CTI platform, we identified several critical objectives to ensure it effectively meets current cybersecurity needs. These objectives span the key areas of data acquisition, data processing, and user interaction, each vital to ensuring the platform's effectiveness and usability. Additionally, modularity is a global design principle, enabling seamless integration of the platform with existing and legacy cybersecurity systems while ensuring flexibility for future extensions.

Regarding data acquisition, we aim for comprehensive, automated, and transparent collection of threat intelligence. Our platform primarily leverages data from the CTA, as our active membership grants us access to over 11 million new threat observations monthly. Nonetheless, we have designed the system to remain open and adaptable, supporting integration with other threat intelligence sources and platforms that utilize the STIX standard. This flexibility ensures our solution remains relevant and valuable across diverse cybersecurity environments and evolving data landscapes.

In terms of data processing, the platform must provide robust real-time processing capabilities to quickly identify and respond to emerging threats. The system employs advanced analytical techniques, including artificial intelligence-driven

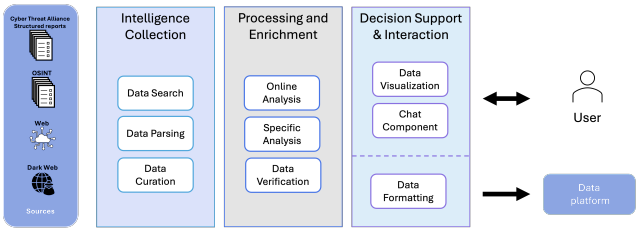


Figure 1. High level design of platform. It includes three different layers: Intelligence Collection, Processing and Enrichment and Decision Support & Interaction

models and predictive analytics, to proactively detect patterns indicative of cyber threats and vulnerabilities. Scalability further ensures sustained operational efficiency even as data volume and complexity grow significantly over time.

For user interaction, our objective is to deliver an intuitive and user-friendly interface tailored for both cybersecurity experts and general IT professionals. The platform emphasizes simplicity and ease of use, allowing users to efficiently access, interpret, and act upon threat intelligence without requiring deep technical expertise. In addition, the system is designed to handle on-demand queries, enabling users to request specific information as needed. This includes the capability to autonomously access new external data sources when necessary to fulfill the user's operational requirements, ensuring that intelligence delivery is both dynamic and context-aware.

B. General Design and Architecture

The platform is structured following a modular three-layer architecture designed to streamline the full lifecycle of CTI processing. The model prioritizes scalability, interoperability, and automation, ensuring that the system can handle large volumes of threat data while maintaining flexibility for future enhancements. The three layers—Intelligence Collection, Processing and Enrichment, and Decision Support & Interaction—work together to transform raw data streams into actionable insights for cybersecurity teams.

Fig. 1 illustrates this layered architecture and highlights the main components operating within each layer.

The first layer, **Intelligence Collection**, is dedicated to automatically acquiring and parsing data from various sources, including external feeds, proprietary datasets, and platforms like the CTA. This layer performs both data search and data parsing operations to ensure raw threat information is captured efficiently and prepared for downstream analysis. By consolidating information from both structured and unstructured sources, this layer guarantees a steady and transparent flow of relevant threat data.

The second layer, **Processing and Enrichment**, is responsible for transforming the ingested data into structured, actionable intelligence. This layer incorporates modules for online analysis, specific contextual analysis, and data verification. Leveraging advanced AI-driven techniques, this processing tier focuses on entity extraction, pattern recognition, anomaly detection, and contextual enrichment of the threat data. The processed information is then prepared for end-user consumption with high accuracy and confidence levels.

The third layer, **Decision Support & Interaction**, is the interface between the platform and the end users. It provides critical components such as data visualization dashboards and a chat-based query interface that allows users to request specific information on demand. A key feature of this layer is its ability to autonomously gather additional external data when necessary to satisfy complex user queries. Additionally, this layer includes a data formatting module that ensures outputs are compatible with external CTI platforms and reporting tools.

C. Specific implementation and platform Components

This section provides a detailed overview of the specific implementation of our CTI platform, outlining the key components already developed and integrated into the system. The platform has been built with modularity and scalability in mind, with each component designed to fulfill a specialized role within the broader architecture. Fig. 2 illustrates the current implementation, showcasing how the components interconnect across the three main layers: intelligence collection, processing and enrichment, and decision support & interaction. The following subsections will describe each of these components and their contribution to the end-to-end workflow of threat intelligence generation and delivery.

1) Intelligence Collection: The intelligence collection layer integrates a suite of specialized components that automate the acquisition and initial processing of data from a variety of sources. The first group of components focuses on parsing information. A key module is the STIX Parser, which efficiently ingests large volumes of structured threat intelligence data. This parser is optimized to process several gigabytes of information in a single operation, streamlining the acquisition of reports such as those shared by the CTA.

In addition to structured data, we have implemented multiple parsers based on LLMs to handle unstructured data sources. These LLM-based parsers are capable of extracting key information from a wide variety of sources, including websites, cybersecurity reports, and social media platforms. By leveraging advanced natural language understanding, these components can distill valuable intelligence from diverse and heterogeneous datasets.

An additional genAI agent has been developed to autonomously navigate Telegram channels. This agent actively monitors hacker communities and cybercrime forums, collecting intelligence on emerging cybercrime-as-a-service trends. It dynamically follows links shared within these channels to identify and collect relevant discussions or leaked data, feeding the acquired insights back into the platform.

Following the acquisition, the collected information undergoes a first curation phase. This is achieved through components leveraging genAI techniques to extract critical elements such as IoCs, Advanced Persistent Threat (APT) group identifiers, TTPs, malware names and targets from the ingested data. Extracted TTPs are mapped to standardized frameworks, such as MITRE ATT&CK, to enhance the interpretability and operational relevance of the intelligence generated. Moreover, summarization modules have been developed to create concise, human-readable reports that simplify the analysis process and enhance situational awareness for cybersecurity professionals.

2) Processing and Enrichment: The processing and enrichment layer combines graph-based analytics and genAI-driven

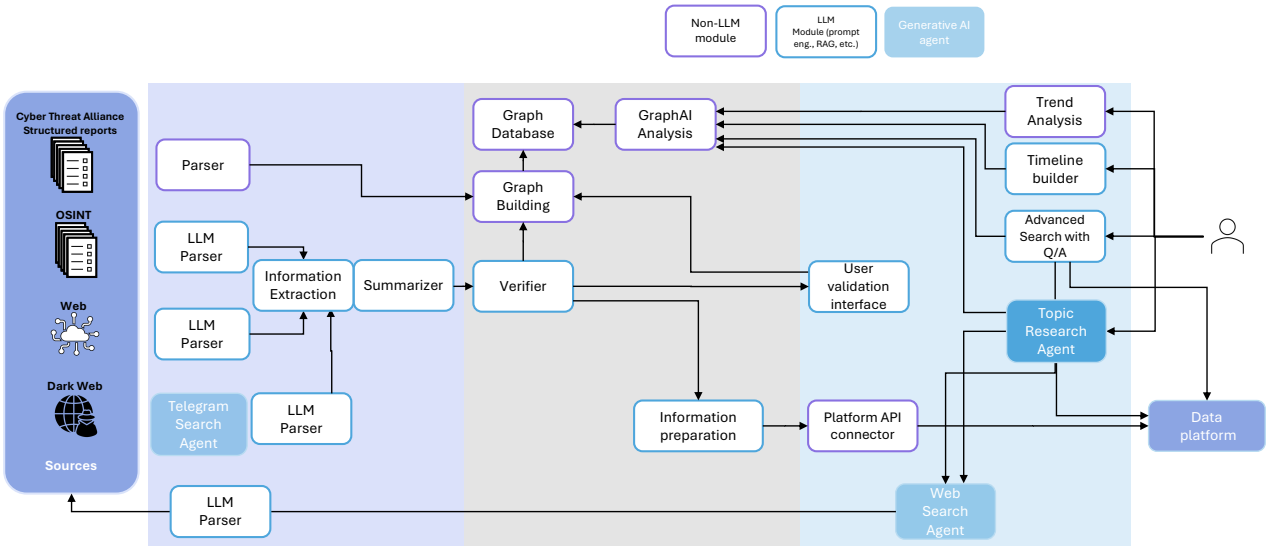


Figure 2. Depiction of the different componets developed for the platform. It includes both genAI powered components and traditional AI components.

components to transform curated intelligence into actionable insights. This phase plays a critical role in structuring, analyzing, and verifying the data collected during the initial acquisition phase.

The process begins with the construction of a knowledge graph using the information acquired and parsed in the previous stage. Leveraging STIX-compliant data, this knowledge graph is built to map entities such as threat actors, vulnerabilities, tactics, and IoCs. The graph-building component also focuses on creating focused subgraphs by identifying key neighborhoods or communities within the larger graph. These subgraphs allow for a more targeted and efficient analysis of specific sectors of the threat landscape.

Subsequently, the GraphAI-based modules analyze the knowledge graph to uncover hidden relationships and patterns between threat actors, tactics, and incidents. By exploring the interconnected data, these modules are capable of identifying correlations and detecting organized malicious campaigns or emerging attack trends that may not be apparent through conventional methods.

In parallel, the platform incorporates genAI modules designed to validate and verify the information extracted from heterogeneous sources. These components help cross-reference and assess the consistency of the data, mitigating the risks of incorporating incomplete or erroneous intelligence that could compromise downstream decision-making.

Finally, this layer includes components responsible for preparing the processed intelligence for integration into the broader platform. These modules ensure that the data are properly structured, enriched and formatted, facilitating its seamless ingestion into the subsequent decision support and interaction layer.

3) Decision Support & Interaction: The decision support and interaction layer provides far more than a simple data presentation interface. It serves as a dynamic environment where users can explore, analyze, and interact with the intelligence generated by the platform. The primary objective of this layer is to enhance operational decision-making through both advanced visualization and intelligent automation tools.

This layer includes a suite of data visualization components that allow users to efficiently navigate the extensive knowledge graph, which stores over one billion STIX objects. Features such as interactive timelines and trend analysis dashboards are provided to give analysts a clear understanding of emerging threats, historical incidents, and evolving attack patterns. These tools empower users to visualize relationships within the data, monitor high-risk areas, and efficiently drill down into specific communities or neighborhoods of interest within the knowledge graph. Fig. 3 shows an example of data visualization.

Beyond visualization, this layer incorporates advanced user-interaction modules powered by two independent genAI agents that collaborate to support user research workflows. The first agent is designed to interpret and process user queries through a chat-like interface. This agent evaluates each request and autonomously determines whether to leverage the platform's internal knowledge base, initiate new analytical operations, or seek external information to fulfill the user's intent. The agent ensures a seamless, conversational experience, enabling cybersecurity experts to interact naturally with the system and obtain tailored responses.

Complementing this, the second agent operates as an autonomous web-search module. When required, this agent executes exploratory searches across multiple search engines, autonomously follows relevant links, and collects external intelligence. The retrieved information is then integrated back into the platform's workflow, enhancing the completeness and contextual relevance of the system's responses. Fig. 4 shows an example of this component.

By combining visualization capabilities with autonomous AI-driven research and retrieval operations, this layer significantly amplifies the platform's effectiveness. It enables cybersecurity analysts to both consume insights and actively engage with the system, facilitating deeper investigations and supporting rapid, informed decision-making in real time.

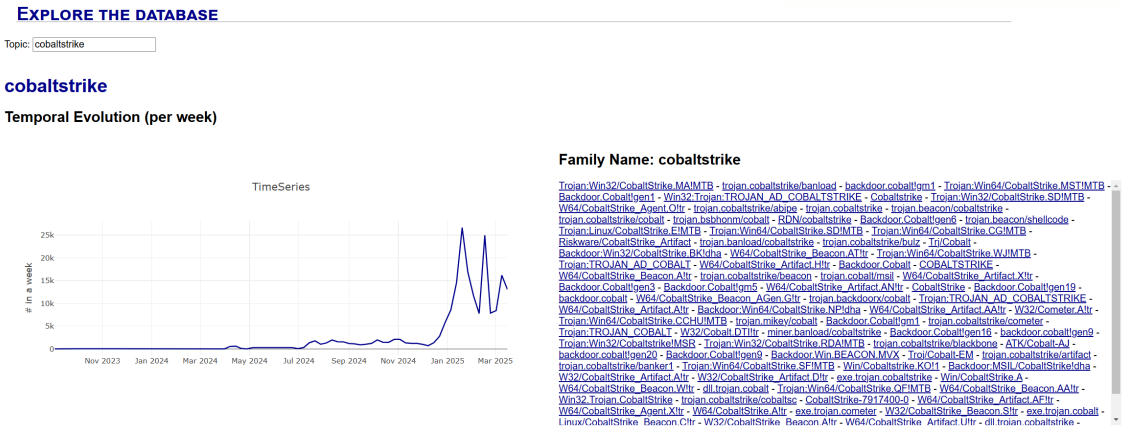


Figure 3. Example of data visualization in the CTI platform. The user queries a specific IoC, vulnerability, or malware, and the interface displays a timeline of its occurrences in the database alongside contextual insights.

D. Detailed Example of a Component: Information extraction

To illustrate how the various components of our platform operate together, we present a detailed example of one of its core modules: the Information Extraction component. This module plays a central role in transforming unstructured or semi-structured threat intelligence into structured, actionable data that can be further analyzed and stored in the knowledge graph. Information extraction is particularly critical for identifying key entities such as IoCs, APT groups, and TTPs, which form the foundation of effective cyber threat analysis.

While this section provides an overview of the module's functionality and integration within the platform, readers interested in an in-depth technical discussion, including evaluation metrics and implementation details, are encouraged to consult our previous work [10] where the full architecture and methodology of the Information Extraction system are described.

The Information Extraction module operates as a key process within the platform, designed to transform unstructured CTI reports into structured and actionable data. The module takes as input a full cyber intelligence report, such as those generated by security vendors, public bulletins, or incident analyses. The system relies on LLMs to automatically extract relevant CTI elements, including IoCs, threat actors, malware families, attack techniques, and targeted sectors.

The extraction pipeline is composed of two stages. Initially, the system applies a pre-processing module that condenses and structures the content of the report to optimize it for LLM consumption. This step is crucial to reduce noise and ensure that only the most pertinent sections of the document are passed to the LLM, while also addressing token limitations.

In the second stage, the LLM is prompted to identify and extract specific CTI entities from the report. For each entity type (e.g., threat actors, tactics, IoCs), the system uses tailored queries to obtain relevant information. Once extracted, the data is passed through a verification phase, where a secondary model ensures that all extracted elements are indeed grounded in the original report.

Finally, the extracted information is automatically mapped to STIX objects and integrated into the platform's knowledge graph, preserving contextual relationships between the entities. This enables subsequent analytical modules to identify patterns and correlations between actors, campaigns, and techniques.

V. Discussion and Future Work

The platform presented in this work demonstrates significant potential in supporting cybersecurity professionals and IT practitioners in navigating the growing complexity of the cyber threat landscape. The system serves as a powerful ally to analysts by automating the collection, enrichment, and presentation of CTI, freeing them from the burden of manually processing vast amounts of raw data. In today's environment, where organizations face thousands of alerts and data points daily, the ability to process, correlate, and present relevant intelligence quickly is essential. Our platform enables users to gain a holistic understanding of emerging threats, actor behaviors, and evolving tactics, which would be impossible to achieve in a timely manner through traditional manual methods.

Beyond streamlining analysis, the platform democratizes access to CTI by offering tools that are accessible to both expert analysts and general IT personnel. The combination of graph-based analytics and genAI modules empowers users to ask sophisticated questions and receive targeted insights, closing the gap between complex technical data and practical decision-making.

From a technical standpoint, the modularity of the system is a crucial advantage. It allows for the seamless integration of new modules and functionalities without disrupting existing workflows. This adaptability ensures that the platform can evolve alongside the rapidly changing threat landscape, as new cyber threats, intelligence sources, and analytical techniques continue to emerge.

Future enhancements may involve deeper integration with evolving versions of the MITRE ATT&CK and D3FEND frameworks to improve threat behavior modeling and defensive technique recommendations.

Looking ahead, one of the most pressing challenges we foresee is ensuring the provenance and reliability of open-source and web-based intelligence. As automated platforms like ours become more common, it is likely that malicious

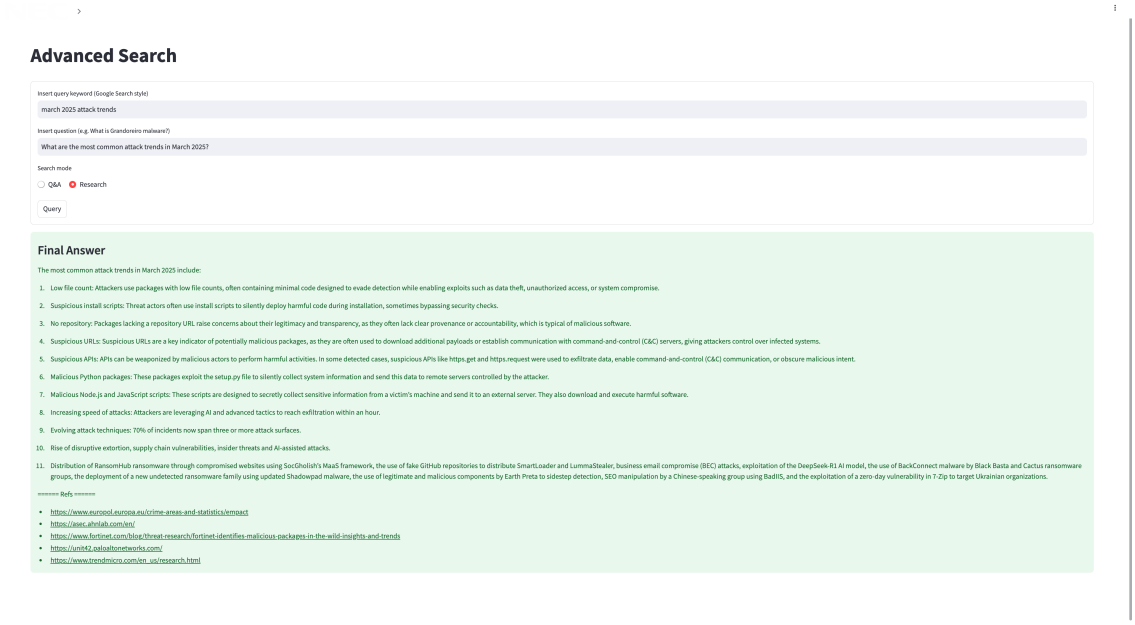


Figure 4. Example of guided search. The user inputs keywords and a specific question, and the system autonomously gathers and analyzes information to generate a relevant answer.

actors will attempt to pollute publicly available datasets with misinformation, disinformation, or decoy content designed to obscure their true activities. This poses significant risks for automated systems that heavily rely on external OSINT sources.

To address this challenge, future iterations of the platform will need to incorporate advanced mechanisms for assessing the credibility, origin, and trustworthiness of external data. This could involve integrating data provenance tracking, source reputation scoring, or even AI models specialized in misinformation detection. By strengthening the system's ability to filter and verify external inputs, we aim to mitigate the risk of adversarial manipulation and maintain high standards of data integrity.

In summary, while the platform already provides robust capabilities for automated threat intelligence analysis and decision support, future research will focus on enhancing resilience against data pollution and improving the platform's adaptability to emerging threats and evolving adversary tactics.

VI. Conclusions

In this paper, we have presented a modular and scalable CTI platform designed to automate the collection, processing, and analysis of large volumes of threat data. The platform enhances the ability of cybersecurity professionals and IT practitioners to navigate complex threat landscapes by providing timely and actionable insights. By combining genAI components with graph-based analytics, the system bridges the gap between data overload and effective decision-making. Our solution reduces manual workload, accelerates threat analysis, and enables users to proactively respond to emerging cyber threats. The architecture's modularity positions the platform for continuous evolution, ensuring its relevance in the face of rapidly evolving cybersecurity challenges.

Acknowledgments

This work has been supported by the Spanish Ministry of Economic Affairs and Digital Transformation through the UNICO Cloud project Privcomp and by the European Commission under the Horizon Europe framework (Empyrean, Grant Agreement No. 101136024) and (Natwork, Grant Agreement No. 101139285, Smart Networks and Services Joint Undertaking).

References

[1] Ömer Aslan et al. "A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions". In: *Electronics* 12.6 (2023). ISSN: 2079-9292. DOI: 10.3390/electronics12061333. URL: https://www.mdpi.com/2079-9292/12/6/1333.

[2] Sergiu Gatlan. *US healthcare provider data breach impacts 1 million patients*. Ed. by Bleeping Computer. [Online; posted 31-January-2025]. Jan. 2025. URL: https://www.bleepingcomputer.com/news/security/data-breach-at-us-healthcare-provider-chc-impacts-1-million-patients/.

[3] Ionut Ilascu. *Orange Group confirms breach after hacker leaks company documents*. Ed. by Bleeping Computer. [Online; posted 25-February-2025]. Feb. 2025. URL: https://www.bleepingcomputer.com/news/security/orange-group-confirms-breach-after-hacker-leaks-company-documents/.

[4] NTT Communications Corporation. *NTT Com Confirms Potential Information Leak due to Unauthorized System Access*. [Online; posted 5-March-2025]. Mar. 2025. URL: https://www.ntt.com/en/about-us/press-releases/news/article/2025/0305_2.html.

[5] Nicolò Boschetti, Nathaniel Gordon, and Gregory Falco. "Space Cybersecurity Lessons Learned from The ViaSat Cyberattack". In: *AIAA Ascend 2022-4380*. Las Vegas, NV, Oct. 2022.

[6] Thomas D. Wagner et al. "Cyber threat intelligence sharing: Survey and research directions". In: *Computers & Security* 87 (2019), p. 101589. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2019.101589. URL: https://www.sciencedirect.com/science/article/pii/S016740481830467X.

[7] Kris Oosthoek and Christian Doerr and. "Cyber Threat Intelligence: A Product Without a Process?" In: *International Journal of Intelligence and CounterIntelligence* 34.2 (2021), pp. 300–315. DOI: 10.1080/08850607.2020.1780062. eprint: https://doi.org/10.1080/08850607.2020.1780062. URL: https://doi.org/10.1080/08850607.2020.1780062.

[8] Mauro Allegretta et al. "Are crowd-sourced CTI datasets ready for supporting anti-cybercrime intelligence?" In: *Computer Networks* 234 (2023), p. 109920.

[9] *The Cyber Threat Alliance*. Accessed: 20-March-2025. URL: https://www.cyberthreatalliance.org/.

[10] Giuseppe Siracusano et al. *Time for aCTIon: Automated Analysis of Cyber Threat Intelligence in the Wild*. 2023. arXiv: 2307.10214 [cs.CR]. URL: https://arxiv.org/abs/2307.10214.

[11] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[12] Feng Xia et al. "Graph Learning: A Survey". In: *IEEE Transactions on Artificial Intelligence* 2.2 (Apr. 2021), pp. 109–127. ISSN: 2691-4581. DOI: 10.1109/TAI.2021.3076021.

[13] Alberto Garcia Duran and Mathias Niepert. "Learning Graph Representations with Embedding Propagation". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/e0688d13958a19e087e123148555e4b4-Paper.pdf.

[14] Shrit Shah and Fatemeh Khoda Parast. *AI-Driven Cyber Threat Intelligence Automation*. 2024. DOI: 10.48550/arXiv.2410.20287.

[15] Romy Fieblinger, Md Tanvirul Alam, and Nidhi Rastogi. *Actionable Cyber Threat Intelligence Using Knowledge Graphs and Large Language Models*. 2024. DOI: 10.1109/EuroSPW61312.2024.00018.

[16] Yongheng Zhang et al. *AttacKG+:Boosting Attack Knowledge Graph Construction with Large Language Models*. 2024. DOI: 10.48550/arXiv.2405.04753.

[17] Filippo Perrina et al. *AGIR: Automating Cyber Threat Intelligence Reporting with Natural Language Generation*. 2023. DOI: 10.1109/BigData59044.2023.10386116.

[18] Yutong Cheng et al. *CTINEXUS: Leveraging Optimized LLM In-Context Learning for Constructing Cybersecurity Knowledge Graphs Under Data Scarcity*. 2024. DOI: 10.48550/arXiv.2410.21060.

[19] Othmane Cherqi et al. *Enhancing Cyber Threat Identification in Open-Source Intelligence Feeds Through an Improved Semi-Supervised Generative Adversarial Learning Approach With Contrastive Learning*. 2023. DOI: 10.1109/ACCESS.2023.3299604.

[20] Rahul Vadisetty and Anand Polamarasetti. *Generative AI for Cyber Threat Simulation and Defense*. 2024. DOI: 10.1109/ICCMA63715.2024.10843938.

[21] Lampis Alevizos and Martijn Dekker. *Towards an AI-Enhanced Cyber Threat Intelligence Processing Pipeline*. 2024. DOI: 10.3390/electronics13112021.

[22] Flavio Ambrosio da Silva. *Navigating the dual-edged sword of generative AI in cybersecurity*. 2025. DOI: 10.34117/bjdv11n1-062.

[23] Cynthia Wagner et al. *MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform*. 2016. DOI: 10.1145/2994539.2994542.

[24] Thanasis Chantzios et al. *The Quest for the Appropriate Cyber-threat Intelligence Sharing Platform*. 2019. DOI: 10.5220/0007978103690376.

[25] S. Bauer et al. *Towards an Evaluation Framework for Threat Intelligence Sharing Platforms*. 2020. DOI: 10.24251/hicss.2020.239.

[26] Alessandra de Melo e Silva et al. *A Methodology to Evaluate Standards and Platforms within Cyber Threat Intelligence*. 2020. DOI: 10.3390/fi12060108.

[27] Yali Gao et al. *HinCTI: A Cyber Threat Intelligence Modeling and Identification System Based on Heterogeneous Information Network*. 2020. DOI: 10.1109/tkde.2020.2987019.

[28] Seulgi Lee et al. *Managing Cyber Threat Intelligence in a Graph Database: Methods of Analyzing Intrusion Sets, Threat Actors, and Campaigns*. 2018. DOI: 10.1109/platcon.2018.8472752.

[29] Eranga Bandara et al. *LUUNU — Blockchain, MISP, Model Cards and Federated Learning Enabled Cyber Threat Intelligence Sharing Platform*. 2022. DOI: 10.23919/ANNSIM55834.2022.9859355.

[30] Saad El Jaouhari and Yaya Etiabi. *FedCTI: Federated Learning and Cyber Threat Intelligence on the Edge for secure IoT Networks*. 2023. DOI: 10.1145/3627050.3627064.

[31] Dimitrios Chatziamanetoglou and K. Rantos. *Blockchain-Based Cyber Threat Intelligence Sharing Using Proof-of-Quality Consensus*. 2023. DOI: 10.1155/2023/3303122.

[32] Prasasthy Balasubramanian et al. *TSTEM: A Cognitive Platform for Collecting Cyber Threat Intelligence in the Wild*. 2024. DOI: 10.48550/arXiv.2402.09973.

[33] Eunsoo Kim et al. *CyTIME: Cyber Threat Intelligence ManagEment framework for automatically generating security rules*. 2018. DOI: 10.1145/3226052.3226056.

[34] Rui Azevedo, Ibéria Medeiros, and A. Bessani. *PURE: Generating Quality Threat Intelligence by Clustering and Correlating OSINT*. 2019. DOI: 10.1109/TrustCom/BigDataSE.2019.00071.

[35] Borce Stojkovski et al. *What's in a Cyber Threat Intelligence sharing platform?: A mixed-methods user experience investigation of MISP*. 2021. DOI: 10.1145/3485832.3488030.

[36] Daniel Fischer et al. *The Role of Threat Intelligence Sharing Platforms in Companies, Public Authorities,*

and Universities: Insights from an Exploratory Global Survey. 2024. DOI: 10.1109/BigData62323.2024.10825242.

Clasificación Automática de Canales Telegram con BERTopic y DeepSeek

F.J. Blanco-Jiménez*, F. González-López*, A. Casares-Andrés†, P. García-Teodoro*
*Network Engineering & Security Group (NESG) – CITIC, Universidad de Granada, España
blancojimenez@correo.ugr.es, {feglez, pgteodor}@ugr.es
†Constella Intelligence, Granada, España
alberto.casares@constellaintelligence.com

***Resumen*—Telegram se ha consolidado como un entorno propicio para actividades cibercriminales debido a su facilidad de uso y moderación limitada. Los enfoques actuales para detectar contenido malicioso en esta plataforma dependen de clasificaciones manuales, lo que restringe su escalabilidad y adaptabilidad a nuevas amenazas. Este trabajo propone una metodología automatizada que combina BERTopic, para identificar dinámicamente tópicos en canales de Telegram asociados a cibercriminalidad, y DeepSeek, un *Large Language Model* (LLM), para etiquetar automáticamente los tópicos detectados. Los resultados, evaluados mediante métricas de clasificación y análisis cualitativo, muestran que este enfoque no solo replica con éxito las clasificaciones de trabajos previos, sino que también identifica nuevas categorías de actividad cibercriminal, como la distribución de *malware*. Este método ofrece una aproximación escalable para la monitorización de amenazas en Telegram, con aplicaciones en la investigación y mitigación del cibercrimen.**

***Index Terms*—Ciberseguridad, Telegram, LLM, NLP, Cibercriminalidad**

Tipo de contribución: *Investigación original*

I. Introducción

En los últimos años, Telegram se ha consolidado como una de las plataformas de mensajería más populares, con más de 900 millones de usuarios activos mensuales [1], [2]. Sin embargo, su creciente adopción, combinada con políticas de moderación más laxas que las de otras plataformas, ha contribuido al aumento de actividades cibercriminales dentro de su ecosistema [3]. A diferencia de los foros tradicionales de la *dark web*, que requieren software especializado para su acceso, los canales públicos de Telegram son fácilmente accesibles, lo que reduce significativamente las barreras de entrada para actores maliciosos y facilita la distribución masiva de contenido ilícito [4].

Estudios recientes, como los realizados por [5], [6] y [7], han identificado cientos de Canales de Actividad Cibercriminal (CAC) en los que se comparten diversos recursos ilegales, incluyendo credenciales de cuentas comprometidas, software y archivos multimedia pirateados, y herramientas diseñadas para realizar ciberataques, como *malware*, *exploits* y estrategias de ingeniería social. Además, una proporción significativa del contenido distribuido en estos canales busca engañar a los propios usuarios mediante técnicas de *phishing* o enlaces maliciosos, lo que representa riesgos considerables para su seguridad [5]. Este escenario evidencia la necesidad de desarrollar métodos automáticos capaces de identificar y categorizar este tipo de contenido de manera eficiente, tanto para la investigación en ciberseguridad como para la mitigación de amenazas en esta plataforma.

Las propuestas actuales para la clasificación de contenido cibercriminal en Telegram suelen depender en gran medida de procesos manuales. Por ejemplo, DarkGram [5] emplea una clasificación manual para etiquetar canales en cinco categorías predefinidas, que luego utiliza para entrenar un modelo BERT en la detección automática de contenido malicioso. Aunque este método logra una precisión cercana al 98 %, presenta limitaciones significativas en términos de escalabilidad y adaptabilidad: (i) requiere una taxonomía fija definida previamente, y (ii) depende de una clasificación manual inicial, lo que dificulta la detección de nuevas tipologías de actividad criminal sin intervención humana.

Para abordar estas limitaciones, este trabajo propone un sistema completamente automatizado para la clasificación de contenido cibercriminal en Telegram, basado en la integración de técnicas de *topic modeling* y modelos de lenguaje avanzados. En primer lugar, se emplea BERTopic [8], un modelo no supervisado de extracción de temas basado en *transformers*, para agrupar las publicaciones de los CAC en tópicos semánticamente coherentes sin intervención manual. En una segunda etapa, se utiliza DeepSeek [9], un *Large Language Model* (LLM), para asignar etiquetas descriptivas a los tópicos identificados, empleando las palabras clave generadas por BERTopic, el nombre del canal y una muestra de publicaciones. Este proceso permite etiquetar los temas de manera dinámica, sin limitarlos a un conjunto predefinido, facilitando así la detección de tendencias emergentes de cibercriminalidad.

Este trabajo demuestra que la combinación de BERTopic y DeepSeek es efectiva para el etiquetado automático de temas, permitiendo la detección dinámica de nuevas tipologías de actividad cibercriminal en Telegram, como la distribución de *malware*.

En este marco de trabajo, la estructura del presente documento es la siguiente. La Sección II revisa trabajos relacionados en esta línea de investigación. En la Sección III se describen los modelos utilizados, el *dataset* empleado, la metodología seguida y las métricas aplicadas. La Sección IV presenta las fases del análisis de canales Telegram y los resultados obtenidos. Finalmente, la Sección V expone las conclusiones y esboza líneas de trabajo futuro.

II. Antecedentes en el Campo

En la actualidad, numerosos estudios han explorado la detección y clasificación de contenido cibercriminal en plataformas de mensajería instantánea, un área de creciente interés debido al uso de estas plataformas para actividades

ilícitas. Entre las primeras aproximaciones automatizadas en Telegram destaca DarkGram [5], que analizó 324 canales vinculados a actividades ilícitas, clasificándolos en cinco categorías predefinidas: credenciales comprometidas, software pirata, contenido multimedia pirata, manipulación de redes sociales y distribución de herramientas *blackhat*. Para establecer estas categorías, los autores realizaron una revisión manual de las descripciones y las diez primeras publicaciones de 4.709 canales recopilados a través de *Telemetr.io* [10], ajustándose a las tipologías de cibercrimen establecidas en el *Federal Bureau of Investigation's Internet Crime Report* (2023) [11]. De este conjunto inicial, identificaron 324 canales relacionados con actividades cibercriminales, los cuales fueron etiquetados manualmente mediante el análisis de diez publicaciones aleatorias por canal, distintas a las empleadas en el filtrado inicial para reducir sesgos. Con este etiquetado, entrenaron un modelo BERT que detectó nuevas publicaciones ilícitas con una precisión promedio del 98 %, lo que permitió analizar 53.605 mensajes y contribuyó a la eliminación de 196 canales cibercriminales.

Aunque DarkGram validó la utilidad de los modelos de lenguaje para la detección de contenido malicioso en Telegram, presenta limitaciones en escalabilidad y adaptabilidad. Su dependencia de una taxonomía predefinida restringe la identificación de tendencias cibercriminales emergentes no contempladas inicialmente, lo que obliga a realizar clasificaciones manuales para incorporar nuevas categorías. Este enfoque resulta poco viable dada la rápida evolución del ecosistema cibercriminal [3]. Asimismo, el etiquetado del conjunto de datos, basado en la revisión manual de tan solo diez publicaciones por canal, compromete su validez como *ground truth* válido, ya que depende de un conjunto limitado de mensajes y está influenciado por la subjetividad de los expertos, quienes también definieron las cinco categorías empleadas. Estos factores limitan la robustez de los resultados, como han señalado estudios previos sobre la fiabilidad de anotaciones manuales en contextos dinámicos [12].

Paralelamente, se han desarrollado iniciativas para recopilar y analizar grandes volúmenes de datos en Telegram. Un ejemplo destacado es TGDataset [6], que reúne información de 120.979 canales públicos y más de 400 millones de mensajes, constituyendo el conjunto de datos más extenso sobre esta plataforma hasta la fecha. A diferencia de investigaciones previas tales como [13], [14] y [15], que se enfocaron en canales específicos por idioma, país o temática, TGDataset ofrece una visión holística del ecosistema. Su construcción partió de 180 *seed channels* seleccionados según su popularidad y categoría, expandiéndose mediante el análisis de mensajes reenviados, una estrategia que maximiza la cobertura pero no garantiza representatividad en dominios específicos como la cibercriminalidad.

Para la clasificación temática, TGDataset emplea *Latent Dirichlet Allocation* (LDA) [16], una técnica de *topic modeling* que infiere temas latentes a partir de la distribución de palabras en el corpus, aplicada a canales en lengua inglesa. Sin embargo, LDA requiere especificar previamente el número de temas, lo que reduce su flexibilidad ante datos dinámicos. TGDataset optimizó este parámetro utilizando la métrica UMass [17], identificando 13 tópicos principales, entre los que se incluyen *carding*, *videogame modding* y contenido para adultos, relevantes para el estudio de actividades cibercriminales. Sin embargo, el enfoque de TGDataset se centró en una clasificación general, sin priorizar específicamente la detección de actividades ilícitas.

LDA, pese a su amplia aceptación en *topic modeling*, tiene limitaciones significativas. Su enfoque basado en frecuencias y co-ocurrencias modela los documentos como una "bolsa de palabras" (*bag-of-words*), ignorando relaciones semánticas profundas entre términos [18]. Además, la necesidad de fijar el número de temas restringe su capacidad para adaptarse a cambios temporales en el contenido, un problema crítico en entornos como Telegram, donde las tendencias evolucionan rápidamente [3]. En contraste, BERTopic [8] aprovecha modelos de lenguaje basados en *transformers* para capturar relaciones semánticas más ricas [19], integrando UMAP [20] para la reducción de dimensionalidad y HDBSCAN [21] para el agrupamiento dinámico de temas. Estas características lo convierten en una herramienta más adecuada para analizar entornos dinámicos y complejos como los canales de cibercrimen en Telegram.

Las técnicas de procesamiento de lenguaje natural (NLP) han probado ser fundamentales para la detección de amenazas en ciberseguridad. Modelos basados en *transformers* han mostrado un rendimiento sobresaliente en la identificación de *spam* y *phishing* [22], así como en la extracción de inteligencia de amenazas a partir de reportes técnicos [23]. Una revisión reciente de Sharma y Singh [24] subraya el potencial de los modelos de lenguaje avanzados para detectar noticias falsas en redes sociales, un desafío análogo a la identificación de contenido malicioso en plataformas de mensajería. Asimismo, sistemas especializados como SecureBERT [23] y CyBERT [25] han demostrado que adaptar modelos de lenguaje al contexto de ciberseguridad mejora significativamente su eficacia, especialmente en dominios con jerga técnica y patrones específicos.

Este trabajo propone una metodología más flexible y escalable que combina técnicas avanzadas de NLP y aprendizaje profundo para superar las limitaciones de enfoques previos. *BERTopic* identifica automáticamente temas en canales maliciosos sin depender de taxonomías fijas, permitiendo la detección de patrones emergentes de cibercrimen. *DeepSeek*, un modelo de lenguaje, refina y etiqueta semánticamente estos temas, optimizando la clasificación con mínima intervención manual. Esta integración supera las limitaciones de enfoques previos, reduciendo la dependencia de la intervención manual y permitiendo una detección dinámica de amenazas en el ecosistema cibercriminal de Telegram.

III. Procesamiento y Clasificación de Canales Telegram

En esta sección se detallan los modelos aquí utilizados para la clasificación temática de canales de Telegram asociados a actividades cibercriminales (CAC), el conjunto de datos empleado, la metodología desarrollada para implementar y validar el enfoque propuesto, y las métricas de evaluación aplicadas. El flujo general del proceso se ilustra en la Fig. 1 y se describe con profundidad en la Sección III-C. Un aspecto central de este estudio, derivado del estado del arte, es que el *ground truth* del conjunto de datos de DarkGram no

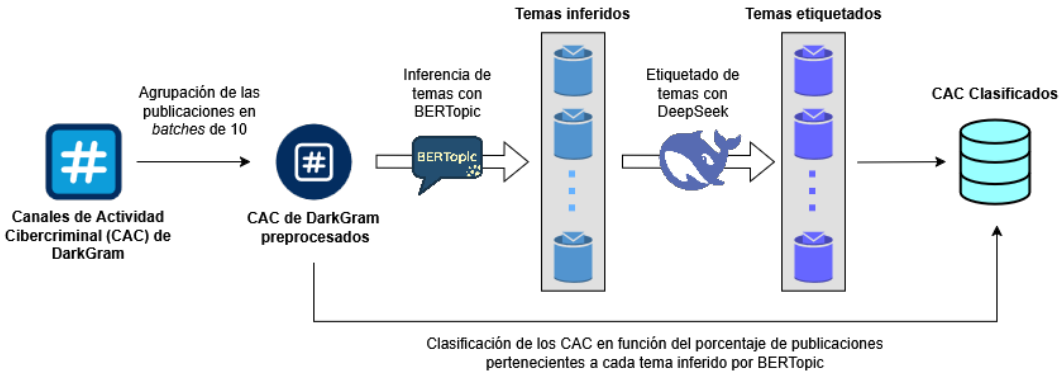


Figura 1. Flujo del proceso de clasificación automática de canales de actividad cibercriminal en Telegram.

puede considerarse absoluto, debido a las limitaciones de su etiquetado manual, lo cual guía tanto el diseño metodológico como la interpretación de los resultados experimentales.

III-A. Modelos empleados

En primer lugar, se emplea BERTopic [8], un modelo no supervisado de extracción de temas basado en *transformers*. Este modelo genera representaciones semánticas de alta calidad mediante *Sentence-BERT* (SBERT) [26], que convierte los mensajes en vectores contextualizados para capturar relaciones semánticas profundas. Posteriormente, aplica reducción de dimensionalidad con UMAP [20], preservando la estructura local de los datos, y, por último, utiliza HDBSCAN [21], para realizar un agrupamiento dinámico de temas sin necesidad de especificar un número fijo de clústeres. Este proceso permite organizar mensajes similares en temas semánticamente coherentes, facilitando la identificación de patrones en los CAC de manera adaptable y escalable.

Seguidamente, se integra DeepSeek [9], un *Large Language Model* (LLM), para asignar etiquetas descriptivas a los temas identificados por BERTopic. DeepSeek, diseñado para tareas de generación y comprensión de texto, refina la clasificación al interpretar el contenido de los temas en función de su contexto semántico. Esta combinación aprovecha la capacidad de BERTopic para descubrir temas emergentes y la precisión de DeepSeek para etiquetarlos de forma interpretable, superando las limitaciones de enfoques basados en taxonomías predefinidas [19].

III-B. Conjunto de datos

Para evaluar la metodología propuesta, se utiliza el conjunto de datos de entrenamiento de DarkGram [5], que comprende 324 canales de Telegram identificados como cibercriminales. Sin embargo, como se ha señalado ya con anterioridad, este conjunto no puede considerarse como *ground truth* absoluto debido a las limitaciones inherentes a su proceso de etiquetado manual, basado en un número reducido de publicaciones y sujeto a sesgos subjetivos [12]. Esta incertidumbre desempeña un papel crucial en la contextualización de los resultados experimentales y en la justificación del enfoque desarrollado. En todo caso, sí constituye un marco de referencia para la experimentación.

La recopilación de los canales se llevó a cabo mediante la plataforma *Telemetr.io* [10], aplicando un filtro inicial para seleccionar canales en lengua inglesa con un mínimo de 10.000 suscriptores, priorizando aquellos con alta actividad y relevancia en el ecosistema de Telegram. De una muestra inicial de 4.709 canales, se identificaron indicios de actividad cibercriminal en las descripciones o en las primeras diez publicaciones, conforme a las tipologías establecidas en el *Internet Crime Report del Federal Bureau of Investigation* (2023) [11]. Tras este filtrado, se seleccionaron 324 canales que cumplían los criterios definidos.

Tabla I
CLASIFICACIÓN EMPLEADA EN DARKGRAM

Categoría	Número de canales
Compromiso de credenciales	69
Software pirata	109
Recursos multimedia pirata	36
Recursos *blackhat*	42
Manipulación de redes sociales	68

Cada canal fue clasificado manualmente en una de las cinco categorías detalladas en la Tabla I, basándose en el análisis de diez publicaciones adicionales por canal, distintas de las usadas en la fase de filtrado con el propósito de reducir sesgos.

El conjunto de datos final consta de 64.801 mensajes recopilados entre febrero y mayo de 2024 mediante la API oficial de Telegram [27]. Además del texto de las publicaciones, se incluye metadatos como el identificador del canal, la fecha de publicación y otros atributos relevantes, como el número de reenvíos o visualizaciones.

III-C. Metodología

A continuación, se describen los pasos seguidos en este estudio, ilustrados en la Fig 1.

III-C1. Preprocesamiento de los datos: Dado que los mensajes en Telegram tienden a ser breves y fragmentados, el preprocesamiento consiste en agrupar las publicaciones de cada canal en *batches* de diez mensajes antes de aplicar BERTopic. Esta estrategia, inspirada en enfoques de *topic modeling* para textos cortos [28], proporciona mayor contexto al algoritmo, mejorando la coherencia y calidad de los temas identificados.

No se aplicaron técnicas adicionales de limpieza de texto, como la eliminación de *stopwords* o la lematización, ya que tanto BERTopic como DeepSeek están diseñados para procesar texto en su forma natural, capturando matices semánticos

sin requerir preprocesamiento extensivo.

III-C2. Categorización de los canales: Tras el preprocesamiento, se emplearon BERTopic y DeepSeek para la clasificación automática de los CAC, integrando una estrategia *N-best* para abordar la incertidumbre inherente al *ground truth* de DarkGram.

Inicialmente, BERTopic identifica temas dentro de los *batches* de diez publicaciones, agrupando *batches* con contenido temáticamente similar sin necesidad de una taxonomía predefinida. Este paso aprovecha la flexibilidad de HDBSCAN para detectar un número variable de temas, adaptándose a la diversidad del contenido cibercriminal. Para garantizar la representatividad, se filtran los resultados eliminando tópicos presentes en menos del 25 % de las publicaciones del canal, estableciendo un máximo de cuatro temas por canal. Este umbral, basado en heurísticas de *topic modeling* [8], considera la posibilidad de que un CAC pueda abarcar diversas actividades ilícitas en lugar de centrarse exclusivamente en una sola. Si ningún tema alcanza el umbral del 25 %, se seleccionan los dos tópicos más representativos para asegurar una cobertura mínima, evitando clasificaciones vacías.

A continuación, DeepSeek asigna etiquetas descriptivas a los temas extraídos, utilizando:

- Los temas generados por BERTopic con una presencia mínima del 25 % en el canal.
- Una muestra aleatoria de 40 publicaciones del canal, proporcionando contexto adicional.
- El nombre del canal, que a menudo refleja su propósito o contenido.

El proceso de categorización se estructura en dos fases complementarias:

- **Fase 1 - Comparación con DarkGram**: DeepSeek emplea las cinco categorías predefinidas de DarkGram mediante un *prompt* basado en *few-shot learning* [29]. Dado que, como se ha señalado, el *ground truth* de DarkGram no se puede considerar absoluto, se implementa una estrategia *N-best*:
  - *1-best*: DeepSeek selecciona la categoría más probable por canal, replicando la estructura de clasificación de DarkGram para una comparación directa.
  - *2-best*: DeepSeek devuelve las dos categorías más probables, reconociendo que las etiquetas manuales pueden ser imprecisas o que las categorías pueden solaparse, lo cual se evalúa posteriormente en las métricas.

  La hipótesis subyacente es que un desempeño moderado en *1-best* no necesariamente indica fallos del modelo, sino la falta de fiabilidad del *ground truth* de DarkGram, mientras que *2-best* ofrece una evaluación más realista al considerar solapamientos temáticos reales y/o errores humanos en la anotación original.
- **Fase 2 - Exploración de nuevas tipologías**: se elimina la restricción de categorías predefinidas, permitiendo a DeepSeek generar etiquetas libres, lo que facilita la detección de patrones emergentes no contemplados en DarkGram.

III-C3. Implementación técnica: La implementación se llevó a cabo en un entorno local equipado con tres GPU NVIDIA RTX 2080, 200 GB de memoria RAM y dos procesadores Intel Xeon de 8 núcleos, garantizando un procesamiento eficiente de los 64.801 mensajes. El flujo de trabajo se desarrolló en Python, utilizando la biblioteca *pandas* para la manipulación de datos, y BERTopic y DeepSeek para la clasificación temática.

BERTopic se configuró con los parámetros predeterminados *(SBERT: all-MiniLM-L6-v2, UMAP: n_neighbors=15, n_components=5, HDBSCAN: min_cluster_size=10)*, sin ajustes específicos de hiperparámetros, optimizados para *datasets* generales.

DeepSeek se ejecutó mediante un *prompt* estructurado que integra los resultados de BERTopic (nombres y palabras clave de temas), una muestra de hasta 40 publicaciones por canal y el nombre del canal. En la primera fase, el *prompt* para *1-best* restringe las etiquetas a las categorías de DarkGram, seleccionando la más probable, mientras que para *2-best* solicita las dos categorías más probables, reflejando posibles solapamientos o errores en el etiquetado manual. En la segunda fase, el *prompt* se adapta para generar etiquetas sin restricciones. Para este proceso se utilizó la versión **R1 de 32B parámetros** de DeepSeek, seleccionada por su disponibilidad gratuita para ejecución local. No obstante, se podría haber empleado otro LLM, como ChatGPT o Gemini.

III-D. Clasificación de Canales

La evaluación del desempeño del modelo se estructuró en dos fases, analizando la replicación de las clasificaciones de DarkGram y la identificación de nuevas tipologías de actividad cibercriminal.

III-D1. Evaluación de la clasificación en las categorías de DarkGram: En esta fase se comparan las etiquetas generadas automáticamente con las etiquetas manuales de DarkGram. Para abordar la incertidumbre generada por no disponer de un *ground truth* absoluto se adoptó una estrategia *N-best*, tal y como se comentó en la Sección III-C2.

Las métricas empleadas incluyen:

- **Precisión *(Accuracy)*:** Porcentaje de canales cuya etiqueta de DarkGram coincide con la categoría asignada en *1-best*, o con al menos una de las dos categorías en *2-best*.

$$\text{Precisión} = \frac{\text{Predicciones correctas}}{\text{Total de instancias}} \times 100$$

- **F1-score por categoría:** Combina *precision* y *recall* para cada una de las cinco categorías de DarkGram, evaluando el desempeño por clase en ambas configuraciones, *1-best* y *2-best*.

$$\text{F1-score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

  Donde las fórmulas para la *precision* y para el *recall* son:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Tabla II
MÉTRICAS DE LA CLASIFICACIÓN EN LAS CATEGORÍAS DE DARKGRAM (*1-best*)

Categoría	Precisión	*F1-score*
Compromiso de credenciales	68,3 %	63,6 %
Software pirata	89,7 %	85,3 %
Recursos multimedia pirata	81 %	87,2 %
Recursos *blackhat*	52,2 %	64,2 %
Manipulación de redes sociales	87,3 %	78,7 %
Promedio	**75,7 %**	**75,8 %**

Tabla III
MÉTRICAS DE LA CLASIFICACIÓN EN LAS CATEGORÍAS DE DARKGRAM (*2-best*)

Categoría	Precisión	*F1-score*
Compromiso de credenciales	73,7 %	77,2 %
Software pirata	94,1 %	91,3 %
Recursos multimedia pirata	89,5 %	93,2 %
Recursos *blackhat*	83 %	87,6 %
Manipulación de redes sociales	96,6 %	89,6 %
Promedio	**87,3 %**	**87,8 %**

- **Matriz de confusión:** Identifica las categorías con mayores tasas de error y analiza confusiones entre clases, destacando posibles solapamientos semánticos o imprecisiones en las etiquetas manuales.

III-D2. Análisis del descubrimiento de nuevas categorías: En esta segunda fase se examinó la capacidad del modelo para detectar tipologías de actividad cibercriminal no incluidas en DarkGram. Para ello, se revisaron las etiquetas generadas por DeepSeek, evaluando su pertinencia como nuevas categorías mediante revisión manual por expertos en ciberseguridad.

IV. EXPERIMENTACIÓN Y RESULTADOS

En esta sección se presentan los resultados obtenidos en las dos fases de evaluación descritas previamente, analizando tanto la capacidad del modelo para replicar la clasificación manual de DarkGram como su potencial para identificar nuevas tipologías de actividad cibercriminal en Telegram. Los experimentos reflejan la incertidumbre inherente al *ground truth* de DarkGram y el solapamiento entre categorías, lo que subraya la necesidad de enfoques flexibles en la clasificación.

IV-A. Primera fase de evaluación

En la primera fase, se compararon las etiquetas generadas automáticamente por BERTopic y DeepSeek con la clasificación manual de DarkGram [5], utilizando las estrategias *1-best* y *2-best* para evaluar el impacto de considerar múltiples etiquetas por canal y abordar la incertidumbre del *ground truth*. Los resultados se presentan en las Tablas II, III y IV.

IV-A1. Resultados de la estrategia 1-best: La Tabla II muestra las métricas de precisión y *F1-score* para la estrategia *1-best*, donde se asigna una sola categoría por canal. El modelo alcanza una precisión promedio del 75,7 % y un *F1-score* promedio del 75,8 %. Categorías como "Software pirata" (89,7 % de precisión) y "Manipulación de redes sociales" (87,3 %) presentan un desempeño destacado, lo que sugiere que estas actividades tienen características bien definidas en los canales analizados. Sin embargo, "Recursos *blackhat*" obtiene una precisión notablemente baja (52,2 %), lo que podría indicar ambigüedad en su definición o un solapamiento con otras categorías. Esta disparidad refleja las limitaciones de restringir la clasificación a una sola etiqueta en un ecosistema tan complejo como Telegram.

La matriz de confusión, mostrada en la Figura 2, confirma estas observaciones al destacar confusiones frecuentes entre "Compromiso de credenciales" y "Manipulación de redes sociales", así como entre "Recursos *blackhat*" y "Compromiso de credenciales". Este solapamiento semántico sugiere que las actividades en estos canales comparten características similares, lo que dificulta una clasificación precisa bajo el enfoque *1-best*.

IV-A2. Resultados de la estrategia 2-best: En la Tabla III, que corresponde a la estrategia *2-best* (donde se asignan las dos categorías más probables), se observa una mejora significativa, la precisión promedio aumenta a 87,3 % y el *F1-score* a 87,8 %. Todas las categorías muestran un incremento en sus métricas, especialmente "Recursos *blackhat*", que pasa de un 52,2 % a un 83 % de precisión. Esta mejora indica que muchos canales clasificados incorrectamente en *1-best* combinan características de múltiples categorías, y que permitir dos etiquetas captura mejor la complejidad del contenido.

Un ejemplo ilustrativo es el de varios canales etiquetados como "Compromiso de credenciales" en DarkGram, pero clasificados como "Recursos *blackhat*" por el modelo de este trabajo. Una verificación manual reveló que estos canales ofrecían tanto credenciales comprometidas como herramientas *blackhat* para su explotación, lo que valida la estrategia *2-best*. Este caso evidencia que el *ground truth* de DarkGram no siempre refleja la naturaleza multifacética de los canales, y que un enfoque más flexible es necesario para una clasificación precisa.

La Tabla IV detalla la distribución de categorías en la estrategia *2-best*, mostrando el número de canales donde cada categoría aparece como principal o secundaria, junto con sus porcentajes. Un resultado clave es la ambigüedad y el solapamiento entre categorías, particularmente con "Recursos *blackhat*", que aparece como categoría secundaria en 119

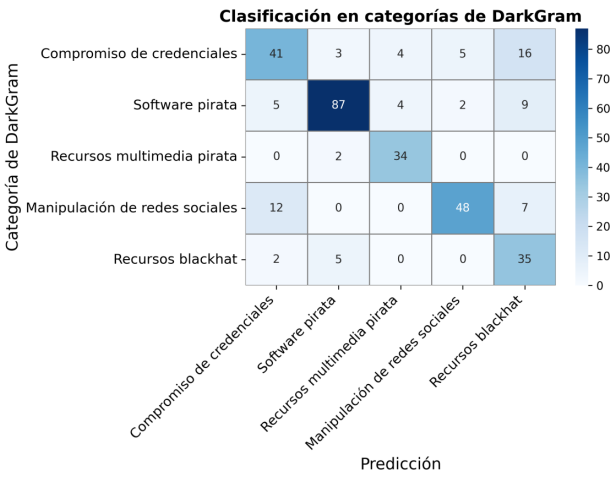


Figura 2. Matriz de confusión de la clasificación en categorías de DarkGram (*1-best*).

Tabla IV
CLASIFICACIONES EN LAS CATEGORÍAS DE DARKGRAM (*2-best*)

Categoría	Número de canales principales	Porcentaje de canales principales	Número de canales secundarios	Porcentaje de presencia secundaria
Compromiso de credenciales	91	28,3 %	49	15,2 %
Software pirata	105	32,7 %	45	14 %
Recursos multimedia pirata	38	11,8 %	34	10,6 %
Recursos *blackhat*	41	12,8 %	119	37 %
Manipulación de redes sociales	46	14,3 %	74	23 %

canales (37 % del total). Este alto porcentaje sugiere que, aunque solo 41 canales (12,8 %) tienen "Recursos *blackhat*" como categoría principal, esta temática está presente como actividad complementaria en una proporción significativa del conjunto de datos.

Por ejemplo, "Software pirata" es la categoría principal más frecuente (105 canales, 32,7 %), pero aparece como secundaria en 45 canales (14 %), lo que podría indicar que algunos canales ofrecen software pirata junto con herramientas *blackhat* o servicios de "Manipulación de redes sociales" (secundaria en 74 canales, 23 %). Este patrón refuerza la idea de que las actividades ilícitas en Telegram suelen mezclarse, y que categorías como "Recursos *blackhat*" suelen ser comunes a una gran cantidad de canales cibercriminales, sea este su principal objetivo o no. Esta ambigüedad justifica la mejora en las métricas de *N-best*, ya que el modelo puede capturar estas relaciones temáticas superpuestas. Ello está acorde con resultados similares en la clasificación de *malware* [30]

IV-B. Segunda fase de evaluación

En la segunda fase, la clasificación libre de DeepSeek permitió identificar nuevas tipologías de actividad cibercriminal en Telegram, ampliando las categorías predefinidas de DarkGram y destacando la coexistencia de múltiples temáticas en un mismo canal. Entre las categorías emergentes se encuentran:

- ***Hindi Dubbed Anime*:** Refleja la distribución de contenido multimedia pirateado con un enfoque regional específico.
- ***Telegram Marketing Services*:** Engloba estrategias de promoción ilícitas dentro de la plataforma. Esta actividad cibercriminal podría considerarse como un caso específico de la categoría "Manipulación de redes sociales".
- ***Mail Access Configurations*:** Asociada al acceso no autorizado a cuentas de correo, podría considerarse como un caso específico de la categoría "Compromiso de credenciales".
- ***Malware*:** Una subcategoría más específica de "Recursos *blackhat*".

El análisis de los resultados revela nuevas categorías emergentes, como "*Hindi Dubbed Anime*", que señala la distribución de contenido multimedia pirateado con un enfoque regional específico, y "*Telegram Marketing Services*", que abarca estrategias de promoción ilícitas en la plataforma. Otras tipologías detectadas incluyen "Mail Access Configurations", asociada al acceso no autorizado a cuentas de correo, y "*Malware*", una tipología mucho más específica de recurso *blackhat*.

Estas nuevas tipologías confirman que los canales de Telegram no se limitan a una sola actividad ilícita, sino que a menudo combinan varias, como se observó en la primera fase con "Recursos *blackhat*". Por ejemplo, un canal de "Software pirata" podría incluir *malware* como parte de su oferta, o uno de "*Telegram Marketing Services*" podría promover servicios relacionados con "Manipulación de redes sociales". Esta naturaleza del ecosistema de Telegram justifica el enfoque flexible adoptado, que permite al modelo adaptarse a patrones emergentes y diversificados en el ecosistema de Telegram.

V. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo presenta un método automatizado para la clasificación de canales de actividad cibercriminal (CAC) en Telegram, integrando técnicas de *topic modeling* con modelos de lenguaje avanzados. Mediante el uso de BERTopic se logra extraer de manera no supervisada los temas predominantes en los canales analizados, mientras que el empleo secuencial de DeepSeek permite asignar etiquetas descriptivas a dichos temas, eliminando la necesidad de intervención manual.

Los resultados confirman la robustez del enfoque en dos frentes. En la comparación con DarkGram, se logra una precisión promedio del 87,3 % con la estrategia *2-best*, superando el 75,7 % de *1-best*, lo que evidencia su capacidad para replicar clasificaciones previas y capturar solapamientos temáticos entre categorías como "Recursos *blackhat*" y "Compromiso de credenciales". Además, la clasificación libre identifica nuevas tipologías, como *Hindi Dubbed Anime*, *Telegram Marketing Services*, *Mail Access Configurations* y *Malware*, destacando la flexibilidad del modelo para detectar patrones emergentes no contemplados en taxonomías fijas.

La asignación de múltiples etiquetas por canal refleja la naturaleza multifacética de los CAC, donde actividades como la piratería de software y la manipulación de redes sociales coexisten frecuentemente. Estos hallazgos subrayan el potencial del método para adaptarse a la evolución del cibercrimen en Telegram, ofreciendo una herramienta valiosa para su monitorización proactiva.

Se identifican varias líneas de mejora para futuras investigaciones:

- Como dirección principal, se propone desarrollar un modelo de detección y clasificación automática basado en BERT, entrenado con las etiquetas generadas en este estudio. Este modelo permitiría analizar publicaciones individuales en tiempo real y asignar etiquetas dinámicas sin depender de la agrupación previa de mensajes en *batches*, mejorando así la granularidad y la inmediatez en la detección de contenido malicioso.
- Asimismo, se explorará la optimización de la metodología mediante el ajuste de hiperparámetros de BERTopic y DeepSeek, así como la incorporación de modelos de

lenguaje especializados en ciberseguridad, como SecureBERT, para mejorar la precisión en la identificación de actividades ilícitas.
- Finalmente, se plantea la posibilidad de ampliar el *dataset* con datos multilingües y multimodales (texto e imágenes) para capturar un espectro más amplio de actividades cibercriminales en Telegram.

Agradecimientos

Este trabajo ha sido desarrollado dentro del "Plan de Recuperación, Transformación y Resiliencia", convenio C025/24 INCIBE-Universidad de Granada, financiado por la Unión Europea (fondos *Next Generation*).

También es parte del proyecto NetSEA-GPT (C-ING-300-UGR23), cofinanciado/a por la Consejería de Universidad, Investigación e Innovación y por la Unión Europea con cargo al Programa FEDER Andalucía 2021-2027.

Referencias

[1] Statista, "Telegram - Datos estadísticos," https://es.statista.com/temas/12679/telegram/, 2024.
[2] Statista, "Aplicaciones de mensajería más populares según el número de usuarios mensuales activos a nivel mundial a julio de 2024," https://es.statista.com/estadisticas/599043/aplicaciones-de-mensajeria-mas-populares-a-nivel-mundial-de/, 2024.
[3] Kaspersky, "La actividad cibercriminal aumentó en Telegram un 53 % en los últimos meses," https://latam.kaspersky.com/about/press-releases/la-actividad-cibercriminal-aumento-en-telegram-un-53-en-los-ultimos-meses-kaspersky, 2024.
[4] A. Radu, "From trust to trade: Analyzing how telegram facilitates cybercriminal activities," Master's thesis, TU Eindhoven, 2024.
[5] S. S. Roy, E. P. Vafa, K. Khanmohammadi, and S. Nilizadeh, "Darkgram: Exploring and mitigating cybercriminal content shared in telegram channels," *arXiv preprint arXiv:2409.14596*, 2024.
[6] M. La Morgia, A. Mei, and A. M. Mongardini, "Tgdataset: a collection of over one hundred thousand telegram channels," *arXiv preprint arXiv:2303.05345*, 2023.
[7] T. Doğaner, "Tracking the evolution of cybercrime on telegram: A scalable tool for continuous monitoring and analysis," Master's thesis, TU Eindhoven, 2024.
[8] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022. [Online]. Available: https://arxiv.org/abs/2203.05794
[9] "Deepseek," https://www.deepseek.com/.
[10] "Telemetr.io," https://telemetr.io.
[11] F. B. of Investigation, "Internet Crime Report," https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf, 2023.
[12] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 1120–1130. [Online]. Available: https://aclanthology.org/N13-1132/
[13] Z. Cao, M. Zheng, Y. Vorobyeva, C. Song, and N. F. Johnson, "Dynamical patterns in individual trajectories toward extremism," *arXiv preprint arXiv:1706.01594*, 2017.
[14] M. La Morgia, A. Mei, F. Sassi, and J. Stefa, "Pump and dumps in the bitcoin era: Real time detection of cryptocurrency market manipulations," in *2020 29th international conference on computer communications and networks (ICCCN)*. IEEE, 2020, pp. 1–9.
[15] A. Urman and S. Katz, "What they do in the shadows: examining the far-right networks on telegram," *Information, communication & society*, vol. 25, no. 7, pp. 904–923, 2022.
[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
[17] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 262–272.
[18] J. Chang, S. Gerrish, C. Wang, J. Boyd-graber, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., vol. 22. Curran Associates, Inc., 2009. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf
[19] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, p. 886498, 2022.
[20] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.
[21] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.
[22] M. Labonne and S. Moran, "Spam-t5: Benchmarking large language models for few-shot email spam detection," *arXiv preprint arXiv:2304.01238*, 2023.
[23] E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "Securebert: A domain-specific language model for cybersecurity," in *International Conference on Security and Privacy in Communication Systems*. Springer, 2022, pp. 39–56.
[24] U. Sharma and J. Singh, "A comprehensive overview of fake news detection on social networks," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 120, 2024.
[25] K. Ameri, M. Hempel, H. Sharif, J. Lopez Jr, and K. Perumalla, "Cybert: Cybersecurity claim classification by fine-tuning the bert language model," *Journal of cybersecurity and privacy*, vol. 1, no. 4, pp. 615–637, 2021.
[26] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
[27] "Official telegram api," https://core.telegram.org/api.
[28] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1445–1456.
[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165
[30] A. A.-G. P. García-Teodoro, J.A. Gómez-Hernández, "Multi-labeling of complex, multi-behavioral malware samples," *Computers & Security*, vol. 121, pp. 1–17, 2022.

Anticipating Adversary Behavior in DevSecOps Scenarios through Large Language Models

Mario Marín-Caballero[1], Miguel Betancourt Alonso[2], Daniel Díaz-López[1,2]
Angel Luis Perales Gómez[3] Pantaleone Nespoli[1] Gregorio Martínez Pérez[1]
[1]Department of Information and Communications Engineering, University of Murcia, 30100, Murcia, Spain
{mario.m.c, danielorlando.diaz, pantaleone.nespoli, gregorio}@um.es
[2]School of Engineering, Science and Technology, Universidad del Rosario, Bogotá, Colombia
{miguels.betancourt, danielo.diaz}@urosario.edu.co
[3]Department of Computers Engineering and Technology, University of Murcia, 30100, Murcia, Spain
angelluis.perales@um.es

***Abstract*—The most valuable asset of any cloud-based organization is data, which is increasingly exposed to sophisticated cyberattacks. Until recently, the implementation of security measures in DevOps environments was often considered optional by many government entities and critical national services operating in the cloud. This includes systems managing sensitive information, such as electoral processes or military operations, which have historically been valuable targets for cybercriminals. Resistance to security implementation is often driven by concerns over losing agility in software development, increasing the risk of accumulated vulnerabilities. Nowadays, patching software is no longer enough; adopting a proactive cyber defense strategy, supported by Artificial Intelligence (AI), is crucial to anticipating and mitigating threats. Thus, this work proposes integrating the Security Chaos Engineering (SCE) methodology with a new LLM-based flow to automate the creation of attack-defense trees that represent adversary behavior and facilitate the construction of SCE experiments based on these graphical models, enabling teams to stay one step ahead of attackers and implement previously unconsidered defenses. Further detailed information about the experiment performed, along with the steps to replicate it, can be found in the following repository: https://github.com/mariomc14/devsecops-adversary-llm.git.**

***Index Terms*—Adversary behaviour, LLMs, attack-defence trees, DevSecOps, Security Chaos Engineering (SCE), threat intelligence, Cyber Situational Awareness (CSA), cyber defence**

Contribution type: *Original research*

I. Introduction

Cyber defense refers to the set of security measures and strategies against cyberattacks [1] and can be implemented either actively or passively. Historically, private and public corporations have mostly relied on passive measures, focusing on patching vulnerabilities. However, active cyber defense allows for a broader perspective and the consideration of a wide range of interaction scenarios between individuals, which has proven to be a more effective strategy [2]. This shift from reactive to proactive approaches has led organizations to seek methodologies that can systematically strengthen their security posture.

One such methodology that supports the implementation of a proactive cyber defense strategy is Security Chaos Engineering (SCE), which improves security and tests the resilience of cloud-deployed architectures through fault injection and controlled experimentation [3]. SCE represents a paradigm shift in how organizations approach security, moving from merely responding to threats toward actively testing defenses before attackers can exploit them. This proactive approach aligns perfectly with modern software development practices that emphasize agility and continuous improvement.

DevOps, which refers to the agile and collaborative integration of development and operations teams [4], embodies these modern practices. When security tasks are integrated from the early stages of development through to software delivery, the practice evolves into DevSecOps. This integration creates a comprehensive environment where security becomes an inherent part of the development lifecycle rather than an afterthought or separate concern, thereby strengthening an organization's overall defensive posture.

An optimal cyber defense strategy in a DevSecOps scenario should integrate vulnerability identification, prioritize threats based on their likelihood of occurrence, explore various attack scenarios while anticipating adversary behavior, and implement automated and effective countermeasures. This holistic approach requires sophisticated tools and methodologies that can model complex attacker-defender dynamics within fast-paced development environments.

Attack-defense trees represent one traditional approach to addressing this need, offering a comprehensive visual framework for cybersecurity analysis by combining offensive and defensive actions in a hierarchical structure. By extending traditional attack trees to include defensive countermeasures, these models enable a dynamic representation of the attacker-defender interplay. This structured approach facilitates improved risk assessment, interdisciplinary collaboration, and strategic decision-making in addressing complex, multi-stage security threats within DevSecOps pipelines.

However, when using these methods (considered traditional approaches for anticipating adversary behavior in cyberspace), their effectiveness is limited by the creativity and knowledge of the security teams implementing them. This limitation creates a significant challenge in maintaining both security and agility

in DevSecOps environments, as human-centered analysis can become a bottleneck in rapid development cycles while still potentially missing novel attack vectors.

In response to these limitations, the state-of-the-art in Large Language Models (LLMs) has emerged as a promising solution by enabling the automation of tasks requiring human language comprehension and generation. Some LLMs, such as RepairLlama [5], which corrects code errors, and Hackmentor [6], which applies fine-tuning to pre-existing LLM models, have been trained directly on cybersecurity data, making them specialized models in this field. Others, like ChatGPT [7], are pre-trained on general data and later adapted using fine-tuning techniques or methods such as in-context learning and chain-of-thought reasoning to tackle complex cybersecurity tasks [8].

Unlike RepairLlama and Hackmentor, the great potential of ChatGPT lies not only in its ability to correct code errors, but also in its outstanding performance across various cybersecurity tasks, such as autonomous vulnerability exploration and the discovery of new threats [9], without requiring a prior transfer learning process, making it readily available for immediate use.

The strategic implementation of LLMs to forecast adversarial behavior patterns (specifically prospective attack procedures, their sequential execution patterns (e.g., branch sequencing within attack-defense tree frameworks), and associated cost-benefit considerations) holds significant potential to mitigate cognitive biases inherent in conventional security testing methodologies within DevSecOps pipelines. Furthermore, such an approach facilitates the systematic exploration of novel attack vectors while refining vulnerability detection and mitigation strategies. Operational integration of such predictive systems within cyber command frameworks could enhance proactive threat modeling, enabling stakeholders to preemptively identify and neutralize emerging threats targeting cloud-based national critical infrastructure assets.

Thus, the main contributions of this paper are summarized as follows:

- A new LLM-based flow for generating attack-defense trees that represent adversary behavior.
- A comparative analysis of attack-defense tree generation is conducted using different general-purpose LLMs, providing the models with a set of prompts within a DevSecOps and cyber defense context to evaluate their ability to generate realistic attack structures with a clear hierarchy aligned with the proposed attack objective.
- The definition of key metrics for assessing the quality of attack-defense trees generated by different models. These metrics ensure that the trees are based on real-world attacks and countermeasures, maintain a clear hierarchical structure, and align with a well-defined attack objective.
- The validation of the proposal through the evaluation of some generated attack-defense trees, and the implementation of a SCE experiment that allow to test the feasibility of the generated paths from the perspective of a realistic adversary.

This paper is structured as follows: Section II explores the LLMs used for cybersecurity and cyber defense purposes. Then, Section III describes our proposed method for generating and evaluating attack-defense trees using general-purpose LLM models. Next, in Section IV, the best-obtained attack-defense trees are used to construct and execute our experiments. Finally, Section V presents the conclusions and analyses potential future research directions to improve our proposal.

II. State of the art

This section describes recent advancements in the use of LLMs for cybersecurity and new frontiers in proactive security analysis. In this sense, the work proposed by Yadong Zhang *et al.* [10] provides a comprehensive survey on leveraging LLMs for strategic reasoning—a complex cognitive ability involving decision-making in multi-agent, dynamic, and uncertain environments. The study systematically explores LLMs' applications, methodologies, and evaluation in strategic scenarios, emphasizing their growing significance in multi-agent interactions and decision-making.

Moreover, Alaeddine Diaf *et al.* [11] present a novel framework for IoT networks that leverages LLMs and Long Short-Term Memory (LSTM) models to predict intrusions by analyzing network packets. The framework uses a fine-tuned Generative Pre-trained Transformer (GPT) model to predict network traffic and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) to evaluate the predicted traffic. An LSTM classifier then identifies malicious packets among these predictions.

Another key research stream is the use of LLMs to discover attack procedures. To this extent, the work proposed by Ying Zhang *et al.* [12] explores the development and deployment of GenTTP, a zero-shot framework that employs LLMs to analyze and automatically generate Tactics, Techniques, and Procedures (TTPs) for interpreted malware specific to Open-Source Software (OSS) ecosystems (e.g., PyPI, NPM). Unlike traditional malware analysis, which relies on manual reverse engineering or dynamic sand-boxing, GenTTP focuses on software supply chain (SCS) attacks, analyzing malicious OSS packages designed to infiltrate systems via deceptive metadata and stealthy code execution.

Furthermore, Lingzhi Wang *et al.* [13] introduce Aurora, a semi-automated Breach and Attack Simulation (BAS) system designed to construct multi-step cyberattack chains efficiently. Aurora leverages LLMs, such as GPT, along with the Planning Domain Definition Language (PDDL), to address challenges in traditional BAS, such as limited action space, expertise requirements, and insufficient attack chain realism. Aurora automates the simulation of full cyberattack chains by analyzing attack tool documentation, threat intelligence reports, and real-world techniques.

The goal of the current paper is to improve the defenses and resilience of systems focused on cyber defense environments. Under this paradigm, Christoforus Yoga Haryanto *et al.* [14] demonstrate how contextualized AI can accelerate strategic decision-making by conducting an automated literature survey

using LLMs. Contextualized AI significantly enhances cyber defense by accelerating threat intelligence synthesis, enabling strategic decision support, improving human-AI collaboration, and facilitating proactive defense posturing.

Also, Johannes F. Loevenich *et al.* [15] focus on the design and training of robust Autonomous Cyber Defense (ACD) agents for use in military networks. It introduces an architecture that combines a hybrid AI model incorporating Multi-Agent Reinforcement Learning (MARL), LLMs, and a rule-based system. These components are organized into blue and red agent teams distributed across network devices. The primary objective is to automate essential cybersecurity functions, including monitoring, detection, and mitigation, thereby enhancing the ability of cybersecurity professionals to safeguard critical military infrastructure.

The previous related works highlight LLMs' transformative potential in cybersecurity and cyber defense. These studies demonstrate LLMs' capabilities in strategic reasoning, IoT intrusion prediction, malware TTP analysis, enhancing BAS systems, synthesizing cybersecurity literature, and building Autonomous Cyber Defense agents. Collectively, they showcase LLMs' versatility in addressing various cybersecurity challenges, from threat detection to automated defense in diverse environments. However, while these contributions demonstrate significant advancements, a crucial gap remains in effectively leveraging LLMs to analyze comprehensive attack procedure databases and integrating this analysis into a SCE methodology. This unexplored approach could enable more sophisticated adversary behavior anticipation, allowing for the proactive design of adaptive security mechanisms that anticipate threats before they materialize, bridging critical gaps in the continuous development lifecycle.

III. Large Language Model as an enabler of adversary anticipation

This section presents the proposed flow for generating attack-defense trees that represent potential adversary behavior, as well as the key metrics defined to evaluate the quality of the trees generated by LLMs.

A. Generating an attack-defense tree

Attack-defense trees provide a structured framework for modeling and analyzing adversarial interactions in cybersecurity scenarios by integrating both offensive and defensive actions into a single hierarchical representation. These trees extend traditional attack trees by incorporating defensive countermeasures, enabling analysts to visualize the dynamic interplay between attackers seeking to exploit vulnerabilities and defenders implementing mitigation. The visual nature of attack-defense trees also enhances interdisciplinary collaboration, enabling technical teams and stakeholders to jointly evaluate security postures and prioritize mitigation against multi-stage threats.

Consequently, in our proposal, the attack-defense trees are used as part of a cyber defense strategy, allowing an organization that uses such trees to evaluate its own critical vulnerabilities and predict adversary behavior patterns according to known attack procedures. As a result, such an organization can pose defense strategies that are adequate considering a cost-benefit analysis and the capacities of the adversary. For the generation of the attack-defense trees, we have provided the LLMs with structured, individual prompts focused on single tasks, rather than a single prompt containing multiple tasks, which offers significant advantages.

LLMs tend to perform better when solving smaller, discrete tasks compared to handling multiple complex tasks simultaneously. This phenomenon can be attributed to the model's inference capabilities, which are constrained by the number of tokens it can process. Another benefit of using multiple individual prompts lies in the model's recall capability. This is particularly important as it enables the LLM to infer attacks and controls that are more contextually aligned with the specific system or infrastructure components being analyzed.

Thus, we propose the flow diagram depicted in Figure 1, which represents the interactions that a cybersecurity analyst should have with an LLM to construct an attack-defense tree. Such a flow will be described next and is composed of the following 6 phases: *Application Security Context, Prompt Context, Insert Prompt, Attack Context, Cosmetic Context, and Expert validation.*

1) Phase 1: Application Security Context: The flow depicted in Figure 1 begins by establishing the context for the LLM model by introducing key concepts such as application security, threat modeling, and attack-defense trees. This step is crucial to ensure that the model remains focused on the relevant subject matter for future tasks and avoids generating content unrelated to these topics. To achieve this, it is possible to pose questions to the LLM about these areas and request summaries of the information.

This approach helps consolidate concise and relevant knowledge within the model's memory, which is inherently limited. Questions like "What is application security?", "What is threat modeling?", and "What is threat modeling using attack trees?" serve as effective starting points.

2) Phase 2: Prompt Context: In the following phase, the LLM model is provided with structured prompts and generates the desired output in the form of an attack-defense tree using DOT format, which is a way to represent these trees as directed graphs using the Graphviz language [16]. The structured prompt format defined for this approach is as follows:

- **System Context**: The context of the system is provided in a specific form, e.g., "The context of the system will be provided by critical mission objectives and operational threat scenarios".
- **Component List**: It contains a list of technologies that will be used to deploy the system, along with the corresponding safeguards for each technology, e.g., "AWS EC2 (monitored by Amazon GuardDuty to detect anomalous accesses)".
- **Attack Goals**: It refers to a phrase that indicates to the LLM that an attack goals is provided, e.g. "Ex-filtrate supply chain information".

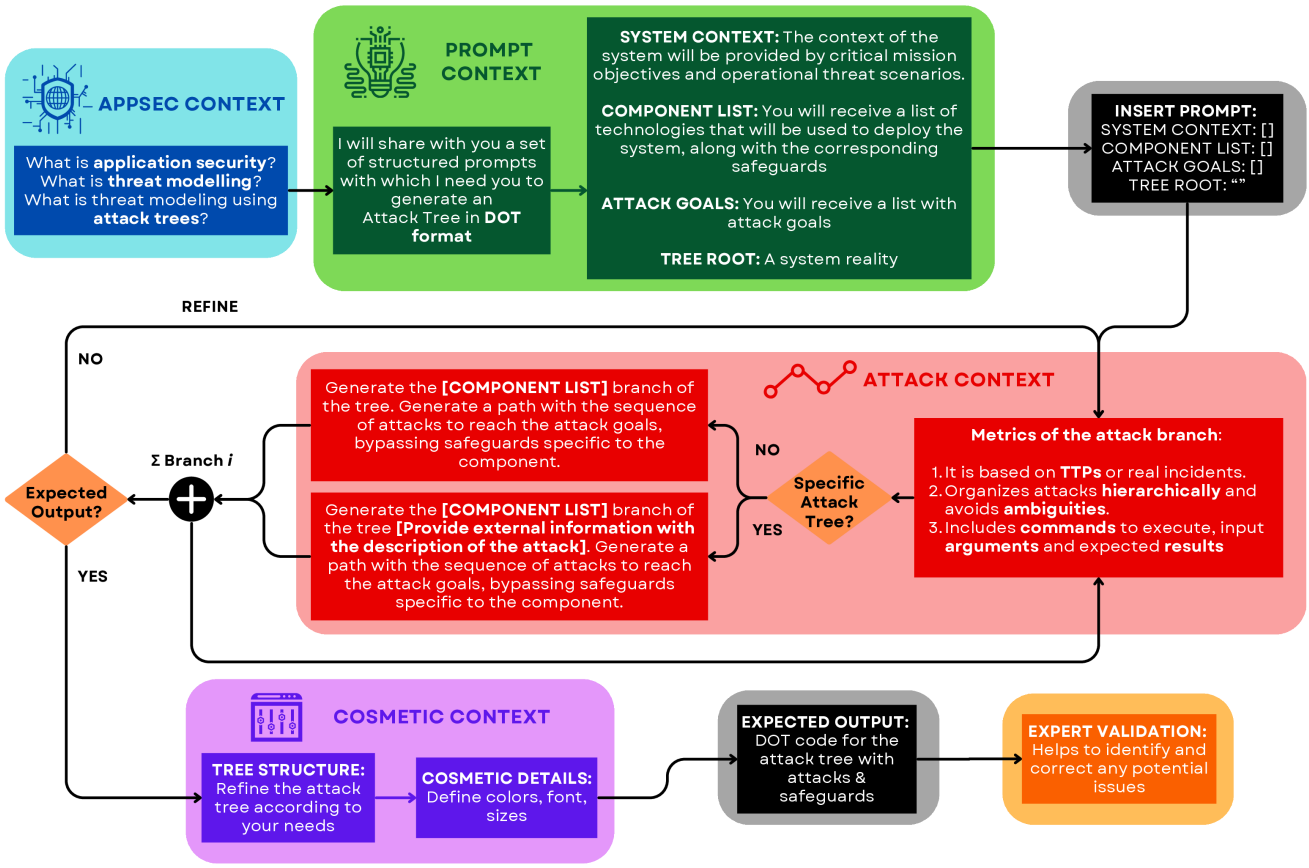


Figure 1: Flow diagram of the construction of an attack-defense tree using LLM

- **Tree Root**: The main node of the tree is specified to the LLM, e.g., "Cloud-based supply chain System".

The decision to use a 4-parameter prompt was based on a series of tests, which revealed that including additional parameters led LLMs to produce meaningless code. It is important to note that the order of the parameters had no impact on the outcomes.

3) Phase 3: Insert Prompt: This phase is critical to the process as it requires summarizing all the knowledge about the system under development. The outcomes at this stage can vary depending on the LLM being used.

4) Phase 4: Attack Context: In order to achieve certain quality in the results produced by LLMs, it becomes necessary to establish a series of metrics. After reading the literature, we can determine that an attack branch should be realistic, meaning it should be based on documented TTPs (Tactics, Techniques, and Procedures) from frameworks such as MITRE ATT&CK or real-world incidents. Additionally, the steps should be organized hierarchically (objectives, steps, and commands) and avoid ambiguities. Finally, the branches should include executable commands, adjustable parameters, e.g., IP addresses, resource names, among others, and expected outcomes.

Thanks to the versatility of the LLM models, it is possible to generate both generalized and specific attack-defense branches. On one hand, generalized attack-defense branches offer a broad overview of possible attacks a system might face. However, it requires a security analyst's expertise to determine whether the inferred attacks are genuinely exploitable within the system. To create a generalized attack-defense branch, the security analyst must prompt the LLM to infer potential attacks based on the system's context, topological components, safeguards, and attack goals.

On the other hand, a specific attack-defense branch provides a more detailed view of the exact sequence of attacks an adversary might follow to achieve their objectives. This approach is particularly useful for security analysts who need precise attack procedures. To create a specific branch, the security analyst must provide the LLM with resources containing a detailed attack, such as one of the attack procedures included in HackTricks [17].

For better results, the security analyst can ask the LLM to generate branches independently for each system component and then combine them into a complete tree. Afterward, it is essential to verify whether the generated attacks align with the intended attack goals. If not, those branches should be refined by iterating the process and including the details needed inside the prompts.

5) Phase 5: Cosmetic Context: Once the security analyst is satisfied with the generated attack-defense tree, he/she can request the LLM to apply cosmetic adjustments, such as modifying the text font, color, size, or node colors, among other visual elements. Additionally, the analyst can instruct the LLM to redefine the tree's structure if necessary. This process may be required in cases where the generated attack-defense tree contains redundant connections between nodes or if the arrangement of parent nodes or attack objectives does not meet the analyst's expectations.

6) Phase 6: Expert Validation: Once the attack-defense tree is generated, it undergoes a thorough validation by a security analyst. If the output does not meet expectations, e.g., if it includes irrelevant attacks, fabricated controls, disconnected nodes, or lacks meaningful relationships between attacks and goals, the analyst can provide feedback. The tree can be returned to the LLM in DOT format with specific instructions for revisions. This expert validation process is crucial for identifying and addressing potential shortcomings, ensuring that the final output is accurate, reliable, and applicable to real-world scenarios.

B. Measuring the quality of an attack-defense tree

In this proposal, these attack-defense trees are intended to support the construction of SCE experiments, and to achieve this objective, they must have a well-defined structure and avoid ambiguities that could delay or even prevent the subsequent validation of attack scenarios aimed at anticipating adversary behavior. To this end, the following metrics are proposed: Attacks based on known TTPs, Ordered attack procedures, and Usable attack procedures.

Attack-defense trees benefit from incorporating the first one as this grounds the model in real-world adversary behaviors, ensuring scenarios align with documented threat intelligence and reducing ambiguity during validation. Ordered attack procedures enforce logical sequences of attack steps, reflecting the dependencies and progression of real-world attacks, which is critical for structuring reproducible experiments and accurately anticipating adversarial workflows. Finally, usable attack procedures prioritize actionable, non-redundant steps that are feasible to implement in controlled experiments, avoiding overly complex or theoretical paths that could hinder practical validation. Together, these metrics enhance the clarity, realism, and operational relevance of attack-defense, enabling systematic testing of defenses against validated attack vectors while minimizing ambiguities in scenario execution.

1) Attacks based on known TTPs: This metric ensures that the tree incorporates real TTPs used in established frameworks such as MITRE ATT&CK or the Cyber Kill Chain framework, which are already well-structured. Thus, minimizing potential hallucinations by the models when generating attack paths.

$$MITRE_{Score} = \frac{\sum_{i=1}^{n} M_i}{n} \times 100 \tag{1}$$

- M_i = Binary value (1 if node i corresponds to a MITRE ATT&CK technique appropriate to the attack being carried out, 0 otherwise)
- n = Total number of attack nodes in the tree

2) Ordered attack procedures: The branches of the tree must be organized hierarchically, starting from a root node that connects the attack nodes and ends in a target node. Additionally, attack paths must be clear and unambiguous, accurately reflecting the process an attacker follows to exploit a vulnerability:

$$Ordered_{Score} = (1 - \frac{N_d + N_{sc}}{n}) \times 100 \tag{2}$$

- N_d = Number of nodes deviated from the original order established in the external source attached to the LLM
- N_{sc} = Number of nodes without children that are not a final node
- n = Total number of attack nodes in the tree

3) Usable attack procedures: Includes commands to execute, input arguments and expected results. Attack and defense nodes must clearly describe what is done (procedure), how it is done (commands and input arguments), and the expected outcome:

$$Usable_{Score} = \frac{\sum_{i=1}^{n}(C_i + I_i + R_i)}{3n} \times 100 \tag{3}$$

- C_i = 1 if the node contains at least one real executable command appropriate to the attack being carried out, 0 otherwise.
- I_i = 1 if the node contains input parameters necessary to execute the command, 0 otherwise.
- R_i = 1 if the node contains descriptive expected results, 0 otherwise.
- n = Total number of attack nodes in the tree.

Finally, we can compute the overall score of the tree with the following equation:

$$Tree_{Score} = \frac{MITRE_{Score} + Ordered_{Score} + Usable_{Score}}{3} \tag{4}$$

IV. Experiments

This section of experiments is composed in the following way: Section IV-A where the cyber defense scenario is established, Section IV-B, where the attack-defense trees are generated, Section IV-C, where one of the branches from the previously generated attack-defense trees is selected and translated into a SCE experiment, and Section IV-D, where the results obtained from the attack generation process are presented and analyzed.

A. Settings

This section demonstrates the practical implementation of the methodology proposed in Section III through a military sector case study motivated by the critical need to model and anticipate adversarial tactics in cyber defense planning. In this case study, a Ministry of Defense (MoD) operates a military logistic information system with a cloud-native approach, designed to get the most benefits of cloud capabilities while also being resilient enough to counter possible sophisticated threat actors. This military logistics system track essential war

fighter supplies across global bases and handles classified data including ammunition, medical kits, and rations, which need to be protected against nation-state cyber espionage tactics. The military logistics system implements a security-first design philosophy which stems from historical analysis of adversaries behavior targeting military logistics systems.

With the aim of defining a specific scenario for the experiments, it is assumed that the military logistics system is deployed in AWS GovCloud and employs three key managed services: EC2 for secure computing, CodeBuild for automated CI/CD pipelines, and CodeGuru for AI-powered code security analysis. This cloud architecture provides a realistic scenario for assessing the capabilities of LLMs in generating attack-defense trees based on the data flow presented in Section III. Building upon this foundation, two specific LLMs were selected for their evaluation in reasoning and cybersecurity applications: GPT-4 and QwQ-32B. These models will serve as the core components for simulating adversary behaviors and creating structured attack-defense models, further enhancing the experiment's relevance and applicability to real-world cybersecurity scenarios. For the experiment, we have used the web versions of both models, available at https://chat.qwen.ai and https://chatgpt.com, respectively.

B. Generating Attack-Defense trees

The attack trees generated when applying the proposed iterative process with both models are shown in Figures 2 and 3. Each branch begins with a node (represented in dark blue) that identifies the specific AWS service being targeted. From there, each branch follows a series of attack nodes (represented in light blue) that detail the specific steps, commands, and techniques used in the attack sequence. Despite the diversity in attack methodologies, all three paths ultimately converge toward the same goal: executing a privilege escalation attack.

- **Branch 1:** In the first branch, i.e., EC2 attack branch, the attacker gains access to EC2 resources, executes a reverse shell to establish a connection with their systems, and steals sensitive credentials from the instance. This allows them to escalate privileges and access additional AWS resources.
- **Branch 2:** In the second branch, i.e., CodeBuild attack branch, the attacker uses valid credentials to set up a compromised CodeBuild project that runs a reverse shell, enabling them to steal credentials from the build environment. They then delete the project to cover their tracks while maintaining unauthorized access to AWS resources.
- **Branch 3:** In the third, i.e., CodeGuru attack branch, the attacker injects malicious code into a repository and manipulates configurations to expose sensitive information during automated code reviews. By extracting credentials from the analysis process, they escalate privileges and gain unauthorized access to sensitive data.

The trees generated by both LLMs were evaluated based on the three key metrics proposed in Section III-B. Overall, as we can appreciate in Table I, a comparison of final scores reveals that QWQ-32B achieved a $Tree_{Score} = 71.60\%$, while GPT-4 scored a $Tree_{Score} = 61.73\%$. In mapping real-world TTPs from the MITRE ATT&CK framework, QwQ-32B shows a more reliable representation with minimal hallucination, compared to GPT-4, which often provides incomplete or incorrect information. Both models excel in creating hierarchical structures for attack procedures, but QwQ-32B offers clearer paths and priorities. Importantly, QwQ-32B outperforms GPT-4 in providing usable attack procedures, offering more detailed commands and expected outcomes. Having all this information considered, this makes QwQ-32B the preferred choice for subsequently detailing an SCE experiment from this tree.

Metric	QwQ-32B	GPT-4
Attacks Based on Known TTPs	Reliably maps real-world TTPs from MITRE ATT&CK with minimal hallucination (22.22%) (4/18 nodes).	Most TTPs are incomplete or wrong, slightly reducing realism (11.11%) (1/9 nodes).
Ordered Attack Procedures	Clear hierarchical structure with unambiguous paths and well-defined priority (100%).	Hierarchical but occasionally introduces minor ambiguities in node relationships (100%).
Usable Attack Procedures	Highly detailed commands, inputs, parameters, and expected outcomes (92.59%) (50/18 nodes).	Provides usable steps but lacks technical depth in some areas (74.07%) (20/9 nodes).

Table I: Comparative table of obtained metrics

C. SCE experiment

The purpose of this experiment is to evaluate the resilience of an AWS environment to potential misuse of permissions. Specifically, the experiment will test whether an attacker can successfully request a Spot Instance, attach a privileged IAM role, and execute a reverse shell script to steal the IAM role credentials, using the permissions $ec2 : RequestSpotInstances$ and $iam : PassRole$.

Based on the goal of the attack tree, i.e., perform a privilege escalation attack, it is possible to define the experiment following the SCE methodology:

- **Observability**: Detection of Spot Instance creation with suspicious user data and anomalous permission usage via AWS GuardDuty.
- **Steady State**: A secure AWS environment where GuardDuty detects privilege escalation, misconfigured permissions are non-exploitable, and security controls function as designed.
- **Hypothesis**: AWS GuardDuty will detect privilege escalation attempts when Spot Instances with misconfigured permissions and malicious user data are launched.

D. Experiment Methodology

This section describes the execution of the SCE Privilege Escalation experiment using a Spot EC2 instance. The automation of the experiment, which replicates the attack from branch #1 of the attack tree shown in Figure 3, culminates in obtaining temporary credentials for the $EC2-CloudWatch-$

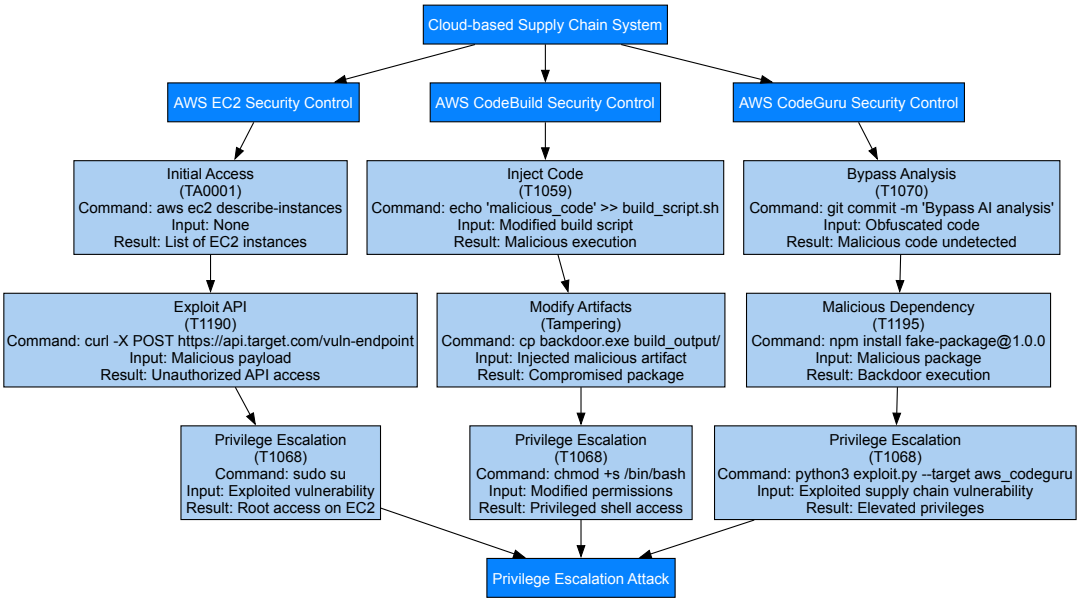


Figure 2: Reduced version of the attack-defense tree generated by GPT-4

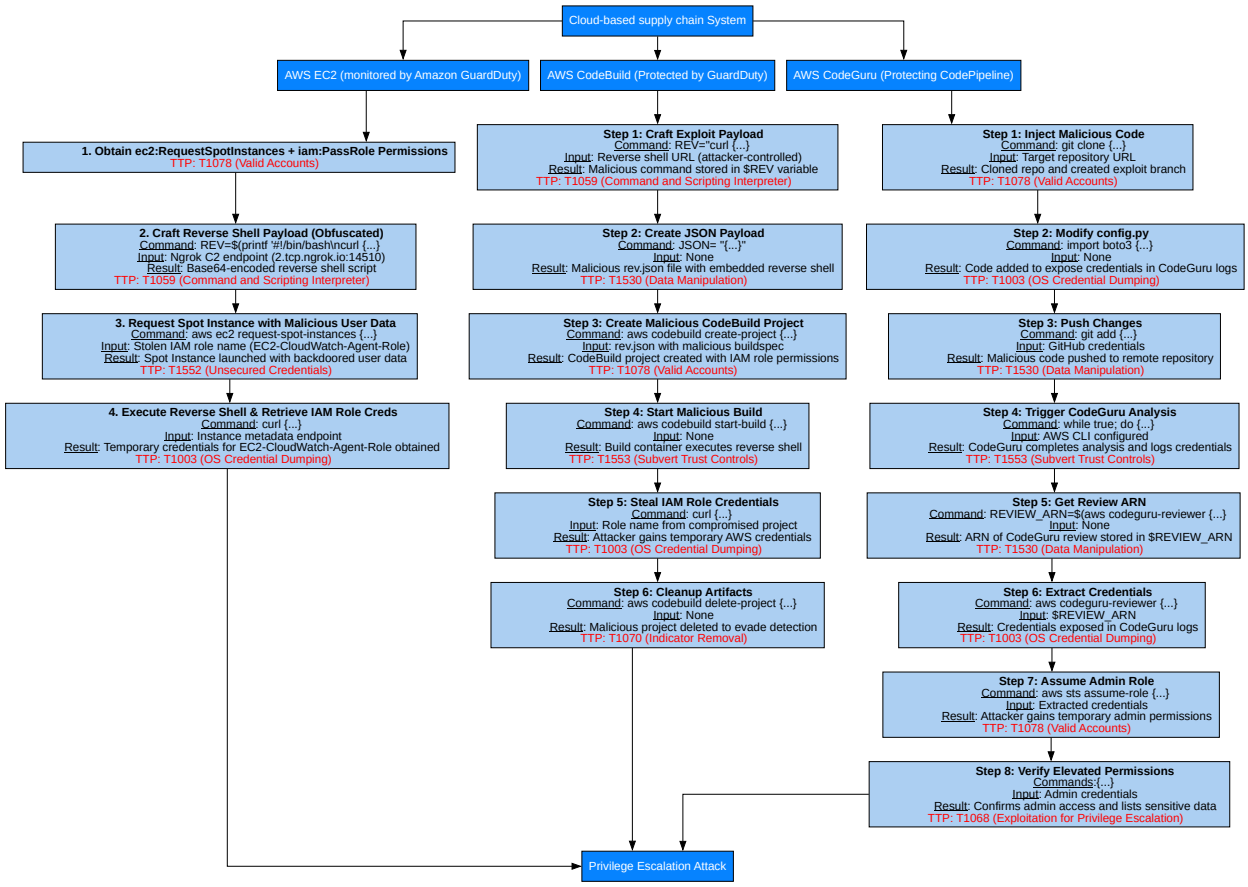


Figure 3: Reduced version of the attack-defense tree generated by QwQ-32B

$Agent - Role$. Further detailed information about the experiment, along with the steps to replicate it, can be found in the following repository: https://github.com/mariomc14/devsecops-adversary-llm.git. Specific actions performed per each stage of the experiment are described next:

- GuardDuty Findings: The initial state of Amazon GuardDuty is verified, ensuring that no prior alerts exist before executing the subsequent phases of the experiment. If no alerts are detected, the execution proceeds with the creation of the Spot EC2 instance.
- Spot EC2 Instance Creation: The command to create the instance with the $EC2 - CloudWatch - Agent - Role$ is executed, incorporating the Base64-encoded string as a payload in the UserData field. During the instance creation process, a connection for reverse shell listening is established in parallel.
- Credential Extraction and Usage: In the third and final stage of the experiment, the obtained credentials are extracted and used by executing the request shown below. These credentials are stored and leveraged to perform actions that the attacker initially did not have permission for, such as listing users.

V. Conclusions and Future Work

In this paper, we proposed an LLM-based flow for generating attack-defense trees that represent adversary behavior. We defined key metrics to evaluate the quality of the generated attack-defense trees. Based on the generated diagrams, we constructed an SCE experiment, relying on one of the previously selected branches of the attack-defense tree. Subsequently, we executed the SCE experiment in ChaosXploit, achieving privilege escalation in an EC2 Spot instance, as part of a military supply chain system in a cyber defense scenario. The results demonstrated the capability of the attack-defense trees generated by our proposed flow to anticipate adversary behavior in exploiting pre-existing system vulnerabilities, and its utility in a cyber defense strategy.

Future work will explore extending this framework to support automated countermeasure recommendations to enhance its practicality for anticipating adversary behavior in real-world scenarios.

Acknowledgment

This work has been co-funded by the European Union (project ECYSAP EYE). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Defence Fund. Neither the European Union nor the granting authority can be held responsible for them. This work has also been partially supported by MCIN/AEI/10.13039/501100011033 NextGeneration EU/PRTR, UE, under Grant TED2021-129300B-I00, by MCIN/AEI/10.13039/501100011033/FEDER, UE, under Grant PID2021-122466OB-I00, by the Spanish National Institute of Cybersecurity (INCIBE) by the Recovery, Transformation and Resilience Plan, Next Generation EU under the strategic project DEFENDER, by the CyberDataLab (Cybersecurity and Data Science Laboratory) at the University of Murcia (Spain), and the School of Engineering, Science and Technology at the University of Rosario (Colombia).

References

[1] J. Srinivas, A. K. Das, and N. Kumar, "Government regulations in cyber security: Framework, standards and recommendations," *Future Generation Computer Systems*, vol. 92, pp. 178–188, 2019.

[2] C. Serban, A. Baluta, D. Bucerzan, D. Gavrilut, A. Sobolev, and S. Oprea, "Malware classification meets vision transformation: A survey," *arXiv preprint arXiv:2304.01142*, 2023.

[3] K. A. Torkura, M. I. H. Sukmana, F. Cheng, and C. Meinel, "Cloudstrike: Chaos engineering for security and resiliency in cloud infrastructure," *IEEE Access*, vol. 8, pp. 123 044–123 060, 2020.

[4] C. Ebert, G. Gallardo, J. Hernantes, and N. Serrano, "Devops," *IEEE Software*, vol. 33, no. 3, pp. 94–100, 2016.

[5] A. Silva, S. Fang, and M. Monperrus, "RepairLLaMA: Efficient representations and fine-tuned adapters for program repair," *arXiv preprint arXiv:2312.15698*, 2023. [Online]. Available: https://arxiv.org/abs/2312.15698

[6] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu, and D. Meng, "When llms meet cybersecurity: A systematic literature review," *arXiv preprint arXiv:2405.03644*, 2024. [Online]. Available: http://arxiv.org/abs/2405.03644

[7] OpenAI, "Chatgpt: Large language model for conversational ai," https://openai.com/chatgpt, 2024, accessed: 2024-03-19.

[8] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, p. Article 160, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3649506

[9] M. Bedoya, S. Palacios, D. Díaz-López, E. Laverde, and P. Nespoli, "Enhancing devsecops practice with large language models and security chaos engineering," *International Journal of Information Security*, vol. 23, no. 6, pp. 3765–3788, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10207-024-00909-w

[10] Y. Zhang, S. Mao, T. Ge, X. Wang, A. de Wynter, Y. Xia, W. Wu, T. Song, M. Lan, and F. Wei, "Llm as a mastermind: A survey of strategic reasoning with large language models," *arXiv preprint arXiv:2404.01230*, 2024.

[11] A. Diaf, A. A. Korba, N. E. Karabadji, and Y. Ghamri-Doudane, "Beyond detection: Leveraging large language models for cyber attack prediction in iot networks," in *2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*. IEEE, 2024, pp. 117–123.

[12] Y. Zhang, X. Zhou, H. Wen, W. Niu, J. Liu, H. Wang, and Q. Li, "Tactics, techniques, and procedures (ttps) in interpreted malware: A zero-shot generation with large language models," *arXiv preprint arXiv:2407.08532*, 2024.

[13] L. Wang, Z. Li, Z. Guo, Y. Jiang, K. Jung, K. Thiagarajan, J. Wang, Z. Wang, E. Wei, X. Shen, and Y. Chen, "From sands to mansions: Simulating full attack chain with llm-organized knowledge," 2024. [Online]. Available: https://arxiv.org/abs/2407.16928

[14] C. Y. Haryanto, A. M. Elvira, T. D. Nguyen, M. H. Vu, Y. Hartanto, E. Lomempow, and A. Arakala, "Contextualized ai for cyber defense: An automated survey using llms," *arXiv preprint arXiv:2409.13524*, 2024.

[15] J. F. Loevenich, E. Adler, A. Bécue, A. Velazquez, K. Wrona, V. Boshnakov, J. Falkcrona, N. Nordbotten, O. L. Worthington, J. Röning, R. Rigolin, and F. Lopes, "Training autonomous cyber defense agents: Challenges & opportunities in military networks," in *MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM)*, 2024, pp. 158–163.

[16] "Graphviz," accessed: May 5, 2025. [Online]. Available: https://graphviz.org

[17] "Hacktricks," accessed: March 10, 2025. [Online]. Available: https://book.hacktricks.wiki/en/index.html

Leveraging LLMs for Automated IDS Rule Generation: A Novel Methodology for Securing Industrial Environments

Manez Moreno, Xabier Sáez-de-Cámara, Aitor Urbieta
Ikerlan Technology Research Centre
Basque Research and Technology Alliance (BRTA)
Arrasate-Mondragón, Spain
{mmoreno, xsaezdecamara, aurbieta}@ikerlan.es

Mikel Iturbe
Department of Electronics and Computing
Mondragon Unibertsitatea
Arrasate-Mondragón, Spain
miturbe@mondragon.edu

Abstract: Industrial Control Systems have become increasingly connected under Industry 4.0, raising the risk of sophisticated cyber threats to critical infrastructure. Traditional Intrusion Detection Systems (IDS) that rely on manually crafted static rules struggle to adapt quickly to new attacks, while machine-learning-based detectors face challenges in interpretability and require extensive domain-specific data. This paper proposes a novel framework that leverages Large Language Models (LLMs) to automatically generate IDS rules for Suricata, addressing the gap between static rule-based security and the need for dynamic, adaptive defenses. We integrate advanced LLMs into an automated workflow for rule generation, refinement, and validation. Through experiments, the framework demonstrates that LLM-generated rules can achieve high detection accuracy and low false positive rates while optimized prompt strategies and moderate packet-level details boost rule accuracy and effectiveness. The results highlight the potential of LLMs to enhance cybersecurity in industrial environments by rapidly producing transparent and effective IDS rules.

***Index Terms*— Large Language Models (LLMs), Automated Rule Generation, Intrusion Detection Systems (IDS), Threat Detection**

Type of contribution: Original Research

I. INTRODUCTION

Modern industrial environments increasingly interconnect Operational Technology (OT) with Information Technology (IT) systems, yielding efficiencies but also exposing critical infrastructure to cyber threats. High-profile attacks like *Stuxnet* and *Triton* have shown that breaches in Industrial Control Systems (ICS) can cause physical damage and safety risks. Indeed, a recent ICS-CERT report noted "33.8% of ICS computers were attacked in the first half of 2021," [1] underscoring the severity of the threat landscape. Intrusion Detection Systems (IDS) are a cornerstone of defense in such environments, continuously monitoring network traffic for malicious patterns. However, conventional IDSs depend on static signature rules that are manually written and updated. Crafting and maintaining these rules is labor-intensive and cannot keep pace with evolving attack techniques. Although Machine Learning (ML) and anomaly-detection approaches have been explored to automate ICS threat detection, they often require large training datasets and struggle with transparency and trust – critical factors in industrial settings where explainability and determinism are valued.

Recent advances in LLMs offer a promising alternative for automating IDS rule creation. LLMs like GPT and others have shown the ability to generate structured text based on prompts, suggesting they could be used to synthesize IDS rules from descriptions of network activity. The research gap addressed by this work lies in whether and how LLMs can be harnessed to produce operationally effective IDS rules for industrial networks, bridging the agility of AI with the interpretability of expert-written signatures. In this paper, we introduce a framework that leverages state-of-the-art LLMs to automatically generate Suricata IDS rules from network traffic captures. Unlike prior approaches that either require extensive model training or produce only coarse-grained signatures, our method uses off-the-shelf LLMs guided by carefully designed prompts to produce refined, context-rich rules. We evaluate the framework on multiple realistic ICS attack scenarios to validate that the LLM-generated rules can detect attacks with high accuracy and low false positives. The novelty of our approach is in using LLMs as an automated *IDS rule writer* systematically exploring prompt engineering, model selection, and input detail granularity to optimize rule quality. Through this, we aim to reduce the manual effort in developing IDS content while maintaining the expert knowledge and transparency that rule-based systems provide.

The primary objective of this research is to develop and evaluate an automated IDS rule generation framework using LLMs for securing industrial networks. To address this objective, we pose four key research questions (*RQ*):

RQ1: Effectiveness of LLMs in IDS Rule Generation – *How effectively can LLMs generate accurate and relevant IDS rules based on industrial network traffic data?* This question examines if LLMs can produce rules that not only parse correctly but also match malicious patterns in traffic.

RQ2: Impact of Prompt Engineering – *How do different prompt strategies influence the quality and consistency of IDS rules produced by LLMs?* We explore *zero-shot* prompting vs. *few-shot* and *chain-of-thought* prompts, hypothesizing that more guided prompts may yield more precise but possibly fewer generalizable rules.

RQ3: Appropriate Evaluation Metrics – *What metrics best capture the effectiveness and reliability of LLM-generated IDS rules?* Beyond simple syntax validation, we identify metrics such as the percentage of correctly parsed rules, the incidence of duplicate or redundant rules, and detection performance (*precision*, *recall*, *F1-score* on malicious traffic) to quantitatively evaluate the generated rules' utility.

RQ4: Packet-Level Feature Inclusion – *How does the inclusion of packet-level information —incorporating minimal, intermediate, and comprehensive feature sets— affect the quality and effectiveness of IDS rules generated by LLMs?* Incorporating detailed packet-level features may enhance the precision and contextual relevance of the generated rules. However, it remains unclear how different levels of feature granularity impact rule accuracy and operational robustness.

By answering these RQs, we seek to understand the conditions under which LLMs can serve as effective tools for automated rule generation in industrial cybersecurity and how to best configure prompts and inputs to maximize their performance.

II. BACKGROUND

This section provides context on industrial threats, network analysis, IDS technologies, LLMs in cybersecurity and the simulated environments used in our experiments.

a. Threats in Industrial Environments

Industrial networks (e.g., factory automation, power grids) face unique threats ranging from espionage and sabotage to ransomware. Unlike IT systems, ICS often involve legacy protocols and safety-critical processes, making the impact of cyber attacks potentially severe [2]. Notorious incidents such as the Stuxnet malware targeting uranium enrichment, attacks on Ukraine's power grid, and the Triton malware in petrochemical plants illustrate how adversaries can disrupt or even take control of physical processes. These threats exploit both IT-facing components (e.g., SCADA servers, engineering workstations) and OT devices (PLCs, sensors) to achieve their goals. A major challenge is that many ICS were not originally designed with security in mind, and downtime for patching is limited. This elevates the need for proactive detection of anomalies or known attack signatures before damage is done. Attack vectors include network-borne exploits (malicious packets or commands), unauthorized device access, and lateral movement from corporate networks into OT.

b. Network Traffic Analysis

An essential task in securing ICS is analyzing network traffic for signs of intrusions. Network traffic analysis involves capturing packet data and inspecting it for abnormal patterns or known malicious signatures. In industrial settings, this can be challenging because protocols like Modbus, DNP3, or MMS are less familiar and often proprietary [3]. Tools such as Wireshark provide dissectors for many industrial protocols, allowing security analysts to inspect ICS traffic at the packet level [4]. By examining traffic captures (PCAPs), analysts can identify attack indicators (e.g., a PLC write command outside normal parameters, or malformed protocol sequences). This analysis is foundational for creating IDS rules.

c. Detection Technologies and IDS

Intrusion detection can be broadly categorized into signature-based and anomaly-based methods. Signature-based IDS (like Suricata or Snort) use a database of known attack patterns (rules) to flag matching traffic. They are effective for known threats and yield low false alarms when rules are well-crafted, but they require continuous updates to cover new threats. Suricata is an open-source IDS/IPS that uses rules with a rich syntax to match packet headers, payload content, and flow context [5]. An example rule might look for a specific Modbus function code with a dangerous value in the packet payload. Human experts typically write such rules after analyzing threats. In contrast, anomaly-based IDS utilize machine learning to model normal network behavior and alert on deviations. These can potentially detect novel attacks but often suffer from higher false positive rates and lack clear explanations for alerts. Given the industrial need for deterministic and explainable defenses, signature-based detection remains very relevant. Suricata rules are transparent (explicitly stating what pattern triggers an alert) and can be reviewed by engineers.

d. Simulated Environments

Conducting live cyberattacks on operational industrial systems is infeasible due to safety and availability concerns. Therefore, simulated ICS environments are leveraged to generate realistic attack data in a safe setting. In this work we use the open-source framework ICSSIM [6] to create a virtual testbed that mirrors a plant network. ICSSIM enables researchers to emulate ICS components (PLCs, HMIs, etc.) and their network communications. A specialized simulator for the IEC-61850 protocol (substation automation) called IEC61850openserver [7] was also utilized for generating MMS traffic in one scenario. By using these simulated environments, the background traffic and protocols in the PCAPs are realistic for a factory setting is ensured (including normal operations like sensor readings and control commands).

e. LLMs in Cybersecurity

LLMs have exhibited remarkable ability to generate coherent text and code from prompts. In cybersecurity, researchers have begun exploring LLMs for tasks such as code security auditing, malware description, and even writing simple detection rules [8]. However, applying LLMs to produce correct and safe security configurations poses challenges. One major issue is hallucination, an LLM might output plausible-looking but incorrect or even syntactically invalid content if the prompt is not precise. Ensuring that the model's output conforms strictly to Suricata's rule syntax is non-trivial. Another challenge is the context limitation: LLMs have a token limit, so providing all details of network traffic may exceed what the model can process. Additionally, LLMs do not inherently "know" what makes a good IDS rule –they must infer it from examples or instructions- thus, prompt engineering becomes critical. Another limitations comes from the LLMs' non-deterministic nature (the randomness factor can lead to variations in each run). This required running multiple attempts or using temperature controls to get stable results. Security practitioners are cautious about automated decisions; hence our design keeps a human-in-the-loop for final validation of any LLM-generated rule.

III. RELATED WORK

Automating IDS rule generation has long been pursued to reduce the burden on human analysts. Earlier methods applied data mining to network data to derive rules. For example, *Fallahi et al.* [9] propose an approach using decision-rule algorithms (Ripper, C5.0) to automatically generate Snort rules from flow records. Their system successfully detected certain attacks (DoS, brute force) using flow-level features (IP addresses, ports), but it neglected deep packet payload inspection. This limitation meant complex attacks could evade detection. Our work extends this idea by incorporating packet payload content into the rule-generation process, thereby improving precision against sophisticated threats. *Sagala et al.* [10] uses honeypot logs to produce Snort rules focused on attacker and target IPs and ports. While this static log-to-rule mapping automates response to observed malicious IPs, it lacks context and adaptability, resulting in coarse signatures that miss nuanced attacks. In contrast, our LLM-based approach generates more context-rich rules (including content patterns, protocol fields, etc.) to reduce false negatives and false positives.

More recently, researchers have begun exploring LLMs in cybersecurity. *Louro et al.* [8] studied fine-tuning GPT-style models to generate syntactically correct firewall/IDS rules. They found that general-purpose LLMs can be repurposed for rule generation but often require extensive domain-specific fine-tuning for acceptable performance. Our framework avoids dataset-specific fine-tuning by leveraging powerful LLMs via prompt engineering, aiming for immediate usability with pre-trained models. This design prioritizes real-time applicability (using commercial LLM APIs to draft rules on the fly) and demonstrates that even without fine-tuning, careful prompts can yield deployable rules. *Houssel et al.* [11] explored the use of LLMs to explain IDS alerts, aiming to enhance analyst understanding. They noted challenges such as accuracy and computational demands that limit real-time detection capabilities. This research aligns with their view of LLMs as assistive tools but extends it further by combining rule generation and explainability. Specifically, our framework leverages LLMs not only to create actionable Suricata signatures but also provides clear, concise explanations detailing the rationale and purpose behind each generated rule, enhancing transparency and trust.. Finally, *Tudosi et al.* [12] developed an automated system to dynamically generate firewall rules from IDS logs and alerts. This reduced manual work, but the generated firewall rules were largely broad IP/port blocks, and the system depended on known attack signatures to trigger rule creation. We improve upon this by having LLMs generate detailed and proactive rules triggered by recognizing attack patterns in raw traffic, not just reacting to alerts. In summary, prior studies underscore the potential of automation and LLMs in security, but also reveal limitations; lack of deep packet context, need for fine-tuning, or reactive scope. Our framework builds on these insights, using LLMs to produce more precise, context-aware IDS rules and systematically evaluating their efficacy in an ICS context.

IV. PROPOSED FRAMEWORK

Our framework for automated IDS rule generation consists of a pipeline that processes raw network traffic data, interacts with an LLM to produce candidate rules, and validates those rules before deploying them. The workflow comprises several stages:

Traffic Input & Preprocessing: The process begins by analyzing packet captures from the industrial environment to identify traffic segments associated with suspected attacks. Initially, the Zeek tool extracts flow summaries from the main PCAP file. Subsequently, flows related specifically to anomalous activities are isolated based on attacker IP addresses, effectively filtering them from the broader set of normal flows. These isolated anomalous flows are then structured into concise textual representations, clearly describing the malicious activity in a log-like format. This IP-based labeling is justified by the controlled, single-attacker setup of our simulated testbed, where the source IP unequivocally marks malicious flows. We acknowledge this approach may not generalize to stealthy or multi-host threats (APTs, compromised legitimate endpoints) and thus flag it as a limitation and a direction for future work.

LLM Prompting Strategy: The framework employs a structured, iterative prompting strategy involving multiple calls to the LLM:

1. *First Prompt (Initial Rule Generation):* This initial prompt receives anomalous flow details along with essential parameters (e.g., the last rule SID, LLM model, temperature, prompt strategy —*zero-shot*, *few-sho*t, or *chain-of-thought*). It generates preliminary Suricata rules accompanied by natural-language explanations that clearly articulate the rationale and intended detection logic.

2. *Second Prompt (Evaluation):* The second call re-submits the initial anomalous flows and the generated preliminary rules. Its role is to classify each rule as acceptable, needing further refinement, or discardable. Minor corrections for consistency are applied at this stage, and rules requiring detailed packet-level context for refinement are flagged for the third prompt.

3. *Third Prompt (Refinement):* Rules flagged for refinement undergo detailed packet extraction using Tshark filters based on the rule's attributes, capturing the first 500 bytes of relevant payloads. The rule, its explanation, and the extracted packet details are fed back to the LLM. This enriched context enhances rule precision, aligning detection patterns more closely with actual attack characteristics.

   Note: Since payload data is only used at this stage, rules not selected for refinement by the LLM lack payload context.

4. *Fourth Prompt (Parser-based Correction):* Finally, the refined rules are validated using Suricata's parser in test mode. If syntactic or structural errors arise, the affected rules, along with specific parser error messages, are resubmitted for further correction. This iterative correction cycle continues until all rules pass validation, capped at three iterations.

Rule Parsing and Validation: The text returned by the LLM is parsed to extract actual rule lines. Here the framework ensures that only syntactically valid rules are kept. In the proposed implementation, a regex-based parser was applied to extract the Suricata rule and it's explanation from the LLM's output. The framework automatically removes obvious duplicates in the model's output and ensures each rule is unique. Each candidate rule is then tested with Suricata's parser to check for syntax correctness. This step catches any minor format issues. If the model included non-critical errors, a optionally last call to the LLM could be performed to correct itself as stated in *"Fourth Prompt (Parser-based Correction)*".

Automated Rule Evaluation: Validated rules are rigorously evaluated against network traffic to determine their effectiveness in detecting malicious activities while minimizing false positives. The evaluation methodology integrates several stages, including Ground Truth creation, rule application, and detailed metric analysis at both the packet and flow levels.

Ground Truth Creation: A reliable Ground Truth dataset is established by manually inspecting a full network capture (main PCAP) using Wireshark. At least one representative packet from each suspicious flow is manually labeled and exported, forming a reduced PCAP file that captures confirmed malicious traffic. To guarantee that entire malicious flows are represented, the *Community ID* is leveraged to correlate the exported packets with all associated packets in the original PCAP. Consequently, comprehensive malicious flows are identified and extracted, forming a complete and accurate Ground Truth dataset.

Application of Rules and Alert Generation: Once the Ground Truth is created, the validated Suricata rules generated by the LLM are loaded into a local Suricata instance, which analyzes the complete original PCAP.

Suricata produces alerts based on matches between traffic patterns and rules, logging these alerts into structured JSON files for subsequent analysis. This alert analysis classifies outcomes into four categories: *True Positives* (Malicious flows correctly identified), *False Positives* (Benign flows incorrectly flagged), *False Negatives* (Malicious flows not detected) and *True Negatives* (Benign flows correctly ignored).

Evaluation Metrics: Rule effectiveness is quantitatively assessed using standard performance metrics, including precision, recall, accuracy, and F1-score, computed at both packet and flow levels:

- *Packet-level metrics:* Measure the correctness of each packet classification individually, providing granular insights. However, relying solely on packet-level evaluation can lead to biases, particularly with attacks that involve extensive flows (e.g., DoS or brute-force attacks). Missing detection of a single large flow can disproportionately reduce precision and recall, skewing overall results.
- *Flow-level metrics:* Evaluate whether entire communication flows —comprising multiple packets— are correctly classified. Flow-level evaluation is particularly relevant for industrial environments, where malicious activities often manifest as continuous streams of packets within a single flow.

This combined packet and flow-level assessment ensures that rule evaluations accurately represent real-world attack scenarios and prevents large, undetected flows from significantly biasing the metrics.

V. EXPERIMENTS AND RESULTS

A series of experiments were conducted to evaluate the framework across three representative attack scenarios in a controlled ICS environment. The experiments were three:

1. *First Experiment: Single-Prompt vs. Three-Prompt Strategy:* Compares the effectiveness of a single-prompt versus an iterative three-prompt chain for IDS rule generation. The single-prompt method generates rules exclusively based on flow-level information, whereas the three-prompt strategy introduces iterative refinement with packet-level context, potentially improving accuracy and detection performance.
2. *Second Experiment: Model, Prompt Strategy, and Temperature Impact.* Evaluates how variations in LLM models, prompt strategies (zero-shot, few-shot, chain-of-thought), and temperature settings influence rule generation quality. This experiment identifies optimal combinations of parameters by assessing precision, recall, and the number of necessary rule corrections.
3. *Third Experiment: Packet-Level Feature Inclusion.* Analyzes whether varying levels of packet-level detail (minimal, intermediate, comprehensive) provided to the LLM enhance IDS rule accuracy. The experiment aims to determine if richer contextual inputs significantly improve rule effectiveness or introduce unnecessary complexity.

Each scenario represents a distinct attack type(s) recreated in the simulated ICS testbed. Within each scenario, every possible experimental configuration —defined by a distinct set of parameters such as *prompt strategy*, *model choice*, *temperature setting*, and *packet-detail granularity*— is evaluated, with 11 iterations performed for each configuration. For each attack scenario, the corresponding PCAP is collected, all configurations are tested, and the generated rules are then assessed for their ability to detect the threat in that same scenario. Presenting results by scenario highlights how rule performance varies both with attack type and with configuration choice.

Scenario 1: SQL Injection + Modbus PLC Write.

This scenario mimics an attacker who first exploits a vulnerable web interface (SQL Injection) to gain a foothold and then issues unauthorized Modbus/TCP commands to a PLC. The combined attack was executed in the ICSSIM environment.

For the *GPT models*, under the one-prompt strategy —where only flow-level details are provided— GPT-4o-mini and o3-mini achieve 100% parsing accuracy *(without the use of the Suricata's parser feedback and last LLM query for correction)*, while GPT-4o reaches approximately 75%. With the three-prompt chain, which incorporates packet-level details for increased rule precision, GPT-4o-mini's parsing accuracy decreases to around 75%, and both GPT-4o and o3-mini drop to 0%. *Only the* GPT-4o-mini *generates duplicate rules*. When the whole pipeline is executed (the three prompts and the parser's feedback) the *zero-shot* strategy achieved the highest *recall* (~64.7%) and strong overall *F1 scores*, indicating broader detection capabilities but at the cost of slightly reduced syntax *accuracy* (~97.7% rules parsed correctly). Conversely, the few-shot strategy produced the most syntactically accurate rules (~99.7% parse success) but significantly limited detection (recall ~32.4%), creating overly specific rules based on provided examples. The *chain-of-thought* strategy balanced these extremes, offering moderate recall (~51.6%) and high syntax correctness (~97.0% parsed), providing stable and iterative refinement. Incorporating intermediate packet-level features boosts accuracy and recall while avoiding the overfitting seen with all features; minimal features result in overly broad rules that miss attacks.

When using a single prompt, *gemini-2.0-flash* achieves about 76% parsing accuracy, *gemini-2.0-flash-lite* around 65%, and *gemini-1.5-pro* 100%. Shifting to the three-prompt approach, *gemini-2.0-flash drops* to roughly 35%, while *gemini-2.0-flash-lite* and *gemini-1.5-pro* each settle near 50%. Regarding duplicates, the 9% of the generated rules by *Gemini 1.5-pro* model were duplicates, reaching percentages of 28% in some iterations. When executing the whole pipeline, the *chain-of-thought* strategy was optimal for Gemini, achieving the highest recall (~68.7%), outperforming zero-shot (~49.2%) and few-shot (~27.7%) significantly. However, Gemini consistently encountered more parsing errors (~87.6% parsed), indicating lower reliability in syntax compliance. Gemini also generated more duplicates and higher variance in rules created, highlighting greater unpredictability compared to GPT and Claude.

In the case of Claude's models, *Claude-3.5-Haiku* achieves about 50% parsing accuracy, while *Claude-3.7-Sonnet* attains around 65%. Under three prompts, *Claude-3.5-Haiku* drops to 0%, and *Claude-3.7-Sonnet* ranges between 55% and 60%. Neither Claude model produces any duplicate. When executing the whole pipeline, the **few-shot** strategy greatly improved recall (~61.2%) compared to zero-shot (~26.9%), though at the expense of parsing accuracy (~94.6%). Claude performed best syntactically in zero-shot (~98.9% parsed) but detected few attacks. Chain-of-thought prompting provided a good middle ground (recall ~53.3%, parsing ~97.9%), suggesting Claude benefits significantly from additional reasoning or examples to enhance detection without major parsing issues. Performance improve with richer packet-level details; while intermediate features offer a solid balance, including all available details further enhances recall and F1-score compared to minimal configurations.

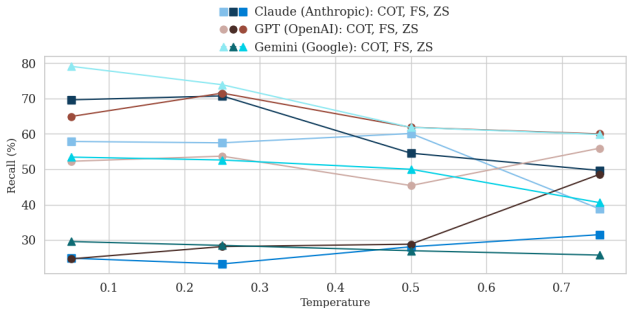


Figure 1: Recall versus Temperature by Family Models for the Scenario 1

Table 1: Average Flow Metrics by Prompt Strategy (aggregated across all models)

Prompt Strategy	*Precision (avg.)*	*Recall (avg.)*	*F1-score*
Zero-shot	~0.89	~0.41	~0.50
Few-shot	~0.94	~0.34	~0.44
CoT	~0.99	~0.44	~0.55

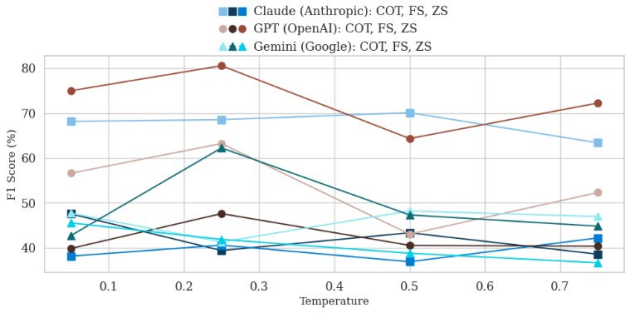


Figure 2: F1-Score versus Temperature for each Model Family and Prompt Strategy (NMAP Scan)

Table 2: Average percentage of rules parsed correctly and Average number of duplicate rules

Model Family	Single-Prompt **Parsed %**	Three-Prompt **Parsed %**	Single-Prompt **Duplicates (avg)**	Three-Prompt **Duplicates (avg)**
GPT	90.9% (±15.3)	65.8% (±37.5)	3.4 (max 102)	0.9 (max 47)
Claude	58.3% (±5.9)	23.4% (±30.2)	0.0	0.0
Gemini	82.2% (±12.2)	41.9% (±34.7)	0.0	0.2 (max 4)

Scenario 2: Nmap Scan of ICS Network.

The third scenario is an active reconnaissance attack: an Nmap port scan sweeps the industrial network to map out open services on PLCs and HMIs. This is a common precursor to targeted attacks. Detecting scans is a classic IDS task; however, this scenario was used to evaluate how LLM-generated rules handle high-frequency events and how they utilize different levels of detail. The network capture contained numerous TCP SYN packets and ARP requests from the scanner.

For the *GPT models*, the non-reasoning variant, *gpt-4o-mini*, proved to be quite conservative with an average *recall* around 28% and an *F1-score* near 0.42, even though in lower temperature *zero-shot* settings it reached roughly 50% *recall* (*F1* ~0.65). In contrast, the reasoning-enhanced *o3-mini* model outperformed its counterpart, achieving about 65% recall and

an F1-score of approximately 0.71 across different temperature settings, while maintaining nearly perfect precision. *Zero-shot* prompting produced concise and consistent rule sets with minimal duplicates, whereas *few-shot* prompting resulted in larger, more variable outputs. The *chain-of-thought* strategy generally led to the highest detection performance —with average *recall* around 44% and an *F1-score* near 0.55— by allowing the model to reason through complex scanning patterns. Additionally, experiments on packet-level feature inclusion indicated that providing *minimal* details (focusing on core features like source/destination IPs, ports, and SYN packet indicators) yielded the best results for GPT models; adding more packet-level information slightly reduced recall (from ~61.4% to ~54.7%) without significant gains in precision.

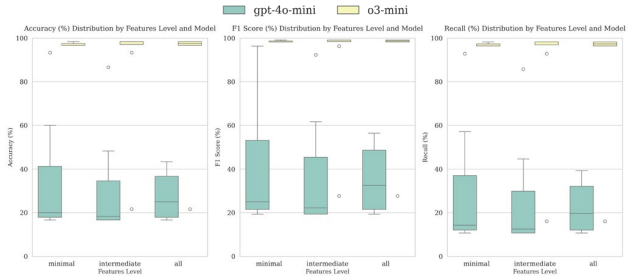


Figure 3: Accuracy, F1-Score and Recall Boxplot for GPT Models for the Scenario 2

Within the *Gemini* family, performance was more variable. The *Gemini-1.5-Pro* model achieved roughly 49% *recall* and an *F1-score* around 0.59 with excellent *precision* (~97.5%), outperforming the newer *Gemini-2.0-Flash* variant, which reached about 40.7% recall and an *F1-score* of 0.49, with *precision* averaging ~85%. The lightweight *Gemini-2.0-Flash-Lite* model struggled considerably, managing only around 15% recall and an F1-score near 0.21, with lower *precision* (~72%). In some cases, *few-shot* prompting marginally improved *recall* (up to ~39.6%) for certain Gemini variants, while *chain-of-thought* prompting boosted detection in low-temperature runs from near-null values to 0.3–0.4; however, these strategies also led to increased duplicate rules and greater variability in outputs.

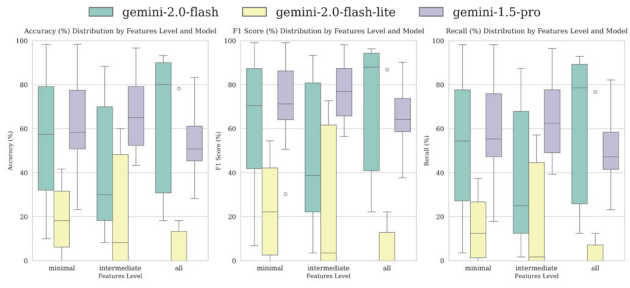


Figure 4: Accuracy, F1-Score and Recall Boxplot for Gemini Models for the Scenario 2

For the Claude models, the advanced *Claude 3.7 Sonnet* clearly outperformed the lighter *Claude 3.5 Haiku* variant, achieving about 47.5% *recall* and an *F1-score* around 0.60 compared to ~26% *recall* and an *F1* of 0.40 for the latter. Both Claude models maintained near-perfect *precision* (virtually 100%) across runs. Under *chain-of-thought* prompting, *recall* for *Claude 3.7 Sonnet* improved dramatically in some cases (up to ~85%), though overall, *few-shot* prompting tended to yield the lowest *recall* (~34%) despite a slight precision advantage. Consistently across the experiments, Claude models benefited most from *minimal* packet-level input; adding excessive packet details introduced noise that significantly reduced both *precision* and *recall*.

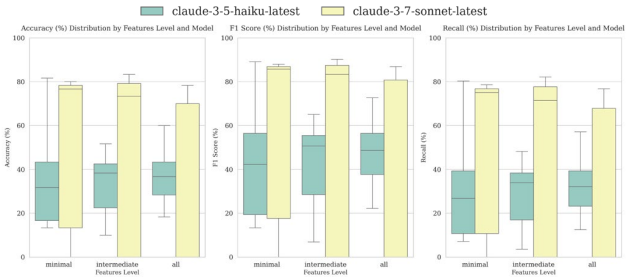


Figure 5: Accuracy, F1-Score and Recall Boxplot for Claude Models for the Scenario 2

VI. DISCUSSION

Our findings demonstrate the viability of LLMs for automating IDS rule creation in industrial contexts, while also revealing important considerations. To revisit the research questions: *RQ1 (Effectiveness)* – The experiments confirmed that state-of-the-art LLMs can generate syntactically correct and operationally effective Suricata rules for diverse ICS threats. In each scenario, at least one LLM produced a high-quality rule or rule set that successfully detected the attack without expert intervention. Notably, certain models stood out: *Claude 3.7* (Anthropic) and *GPT o3-mini* (OpenAI) consistently yielded accurate rules when given clear prompts, detecting even sophisticated multi-step attacks. We observed that Claude models tended to be more deterministic and produced fewer errors, which suggests their training favored reliability – an attractive trait for security applications. Gemini (Google) models generated valuable rules as well, though with higher variability; they sometimes offered very concise solutions and other times overly verbose ones, indicating that iterative prompting or temperature tuning was needed. In summary, *RQ1* is answered affirmatively: LLMs (especially Claude and GPT) can indeed serve as an effective "*virtual analyst*" to write IDS rules, given proper configuration.

For *RQ2 (Prompt Engineering)*, our results highlight that prompt strategy significantly influences outcomes. A *zero-shot* approach was simpler and often yielded a correct basic rule, showing higher initial parse success rates across all model families. However, these rules were sometimes over-generic (catching the obvious pattern but potentially missing subtle variations, or creating redundant rules). Using a multi-step chain-of-thought prompt forced the model to be more thorough and reduced redundant outputs (we saw dramatically fewer duplicate rules with chain-of-thought). The trade-off was that the added complexity could confuse the model's rule formatting, lowering the immediate parsing success. Few-shot prompting (providing examples) proved beneficial for syntax. Models almost never made format mistakes after seeing an example, and correctness was high. However, those examples can bias the model towards copying the specifics of the example, sometimes making rules too narrowly tailored

("overfitting" to the example's pattern). The key insight is that there is no one-size-fits-all prompt: one must balance simplicity for correctness and complexity for completeness. In practice, a hybrid approach may work best: e.g., first use a zero-shot prompt to get a base rule, then use a refined prompt to improve it. Security analysts can then choose which result to deploy. This finding underscores prompt engineering as a crucial skill for applying LLMs in cybersecurity tasks.

Addressing *RQ3 (Evaluation Metrics)*, we found a combination of syntactic and semantic metrics necessary to fully assess LLM-generated rules. Syntactic parsing accuracy (percentage of rules that load without error) was a fundamental metric; in our tests, parsing accuracy above ~90% was needed for a model to be practically useful. All top-tier models met this threshold under optimal prompts, with Claude reaching ~100% in many cases. We also measured the number of duplicate rules and any Suricata parser errors as indicators of output quality. High duplicate counts indicate the model might be "spamming" patterns, which could burden an IDS with redundant checks. Our multi-scenario tests showed that after prompt tuning, duplicates were minimal for most models (often zero or one duplicate at most). Importantly, we introduced flow-level detection metrics (*precision*, *recall*, *F1-score* on malicious vs. benign flows) to directly evaluate if the rules perform their security function. This bridged a gap: a rule that parses correctly isn't useful unless it actually catches attacks and ignores normal traffic. By combining these metrics, we could identify the best configurations. Overall, the chosen metrics effectively captured both the technical correctness and the operational efficacy of generated rules, they can serve as a template for future evaluations of AI-generated security rules.

For *RQ4 (Packet-Level Feature Inclusion)*, the experiments revealed a nuanced but clear pattern: intermediate detail is optimal. When LLMs were given only minimal info, the resulting rules lacked precision or missed specifics (low recall/F1). Conversely, when flooded with all packet bytes or exhaustive details, models occasionally fixated on irrelevant aspects or became inconsistent in formatting, hurting parsing accuracy and sometimes recall. The intermediate approach – providing the most relevant fields and a bit of context– produced the best outcomes. This likely works because it gives the LLM enough context to understand the attack (nuanced behaviors) but not so much that it loses focus or runs out of its attention span. Practically, this means that when using LLMs for rule generation, one should carefully select what input data to include.

Beyond the *RQs*, we observed some model-specific trends worth noting. GPT models emerged as strong overall performers, they often produced effective rules across all scenarios, showing versatility. They did, however, sometimes require the chain-of-thought prompt to reach that effectiveness, indicating they respond well to reasoning tasks. Claude models were remarkably good at maintaining Suricata syntax and providing what felt like professionally written rules. Their deterministic nature (less randomness) is a double-edged sword: great for consistency, but if they misunderstood the prompt, they would consistently do so. We mitigated that by prompt clarity. Gemini models improved significantly with iterative prompting; their first attempt might be messy, but they quickly adapted, which suggests that an interactive loop proposed can harness them effectively. In terms of speed and cost, smaller models (like *GPT-4o-mini* or *Claude 3.5 Haiku*) were faster and cheaper to run, yet still yielded decent rules. This raises an operational point: one could use a cheaper model for quick initial rule drafts and then validate or refine with a more powerful model if needed.

From a security operations perspective, these results indicate that an LLM-driven rule assistant could soon become a reality. An analyst could feed network logs of a new attack into such a system and get a candidate rule, saving precious time during incident response. Our framework's output is human-readable rules, meaning the final decision remains with analysts. They can inspect the AI-generated rule and decide to deploy it, combining AI speed with human judgment. We also note that instructing LLMs to consider negative cases (what not to alert on) by embedding domain knowledge, is useful.

Finally, we must consider limitations. One limitation is token/context size (complex scenarios might exceed the LLM's input limit). We managed this by trimming inputs, but as attacks get larger, this could be problematic (discussed in section VII). Another limitation is that our evaluation was on a controlled set of attacks; real-world traffic might introduce noise that could confuse the models or cause false positives. We did not deeply test the robustness of rules against slight attack variations (beyond the ones the LLM itself considered).

However, given the generally specific nature of the rules, they might need updates if attackers significantly change tactics. Encouragingly, the LLM can simply be re-prompted with new data in such cases. The framework is not yet fully autonomous, we foresee it as a recommendation system for human analysts rather than a black-box rule generator deployed without oversight.

VII. FUTURE LINES

While the framework is effective, there are several avenues to explore to further mature this approach for real-world deployment:

Handling Large or Continuous Data: One immediate area for improvement is dealing with LLM token limitations. Future research can investigate intelligent data segmentation (breaking traffic into chunks that the LLM can process sequentially, perhaps with a rolling context). Techniques like summarizing or clustering similar traffic can reduce input size without losing important information. As newer LLMs with larger context windowsbecome available, the framework should be updated to leverage those, feeding them more complete views of traffic. Maintaining a modular design will allow incorporating such advanced models seamlessly. The ultimate goal is to handle live data streams, potentially by sliding window analysis where the LLM processes recent traffic slices for rule generation suggestions.

Real-time and Autonomous Operation: Currently, a human validates and deploys the LLM-generated rules. Moving towards real-time applicability, we envision integrating this framework with a digital twin or testbed environment for automatic validation. For example, before pushing a rule to a production IDS, it could be tested in a virtual replica of the network (using a simulation) to ensure it catches the intended

malicious behavior and does not over-trigger. This automated testing loop could enable a more autonomous system that generates and verifies rules on the fly. Additionally, research into reducing the need for manual oversight through more robust self-checks is warranted.

Improving Consistency (Reducing Non-determinism): The stochastic nature of LLM outputs means each run might yield a different rule. While we found ways to minimize this (low temperatures, etc.), completely eliminating variability would boost trust. Future work could explore fine-tuning an LLM on cybersecurity data –e.g., training it with lots of example attacks and correct rules– to see if that yields more consistent and domain-optimized outputs. Fine-tuning might also reduce prompt complexity needed, as the model internalizes how to map traffic to rules. Techniques like programmatic chain-of-thought, where the model's reasoning is constrained, could yield repeatable results. We believe achieving higher predictability is crucial for critical infrastructure security adoption, and it's an active area for model improvement.

Robust Output Parsing and Format Handling: We relied on regex parsing to extract rules from the LLM output, which worked but required maintenance (e.g., updating the regex when a model formatted output unexpectedly). Future iterations could employ more sophisticated techniques like using an LLM or parser to interpret the output. Alternatively, incorporating a suricata rule grammar into a parser or using ML-based parsing could make the extraction step more robust to variations. By improving this aspect, the system becomes less brittle and can handle any formatting quirks from future models.

Expanding to Other Security Domains: While our focus was Suricata IDS rules, the concept can extend to other rule-based systems – for example, firewall rules, SIEM alert rules, or even detection code (like Snort dynamic preprocessor or YARA rules for malware). Future work could test LLMs on generating those artifacts, potentially using a similar approach. Each domain has its syntax and semantics, so some adaptation in prompt and validation would be needed.

On-premise and Privacy-Preserving Models: One practical direction is deploying this capability in environments where sending data to cloud APIs is not acceptable (common in critical infrastructure due to privacy). Future lines include evaluating the performance of open-source LLMs that can run on local hardware. If a smaller fine-tuned model can run on an edge server and produce rules nearly as well as those generated by the models tested, that would be a significant win for adoption. Techniques like model distillation or quantization might be applied to compress the knowledge of a large model into a lightweight one that an organization can use internally.

The next steps aim to make the LLM-driven IDS rule generation more scalable, autonomous, and integrated. By overcoming current limitations –token limits, need for human review, output variability– we move closer to an era of self-updating IDS, where new threats are rapidly countered by AI-crafted defenses. Our work provides a foundation, and these future directions outline a path to transition this prototype into a robust, field-ready system. We anticipate that as LLM technology and cybersecurity datasets evolve, the synergy between them will lead to increasingly intelligent and reliable security infrastructure, reducing the window of exposure to new threats and easing the burden on human analysts.

ACKNOWLEDGMENTS

This work has been supported by CRITIC Project Grant PLEC2024-011222 funded by AEI/10.13039/501100011033, FEDER and EU.

BIBLIOGRAPHY

[1] Kaspersky, «Threat landscape for industrial automation systems. Statistics for H1 2021,» 9 September 2021. [Online]. Available: https://ics-cert.kaspersky.com/publications/reports/2021/09/09/threat-landscape-for-industrial-automation-systems-statistics-for-h1-2021/.

[2] R. Masum, «Cyber Security in Smart Manufacturing (Threats, Landscapes Challenges),» *arXiv,* 20 April 2023.

[3] J. F. &. R. K. W. Kurose, «Computer Networking: A Top-Down Approach (8th ed.),» Pearson, 2021.

[4] Wireshark, «Wireshark Documentation,» [Online]. Available: https://www.wireshark.org/docs/.

[5] A. Y. H. L. Y. S. a. L. S. Yan Hu, «A survey of intrusion detection on industrial control systems,» *SAGE,* p. 14, 2018.

[6] A. Dehlaghi, «ICSSIM | GitHub,» [Online]. Available: https://github.com/AlirezaDehlaghi/ICSSIM.

[7] robidev, «iec61850_open_server | GitHub,» [Online]. Available: https://github.com/robidev/iec61850_open_server.

[8] R. A. J. C. C. J. B. F. S. a. P. R. M. I. Bernardo Louro, «Analysis of the Capability and Training of Chat Bots in the Generation of Rules for Firewall or Intrusion Detection Systems,» *International Conference on Availability, Reliability and Security,* p. 7, 2024.

[9] A. S. a. M. Naser Fallahi, «Automated Flow-based Rule Generation for Network Intrusion Detection Systems,» *Iranian Conference on Electrical Engineering (ICEE),* 2016.

[10] A. Sagala, «Automatic SNORT IDS Rule Generation Based on Honeypot Log,» *International Conference on Information Technology and Electrical Engineering (ICITEE),* 2015.

[11] P. S. S. L. a. M. P. Paul R. B. Houssel, «Towards Explainable Network Intrusion Detection using Large Language Models,» *arXiv,* 2024.

[12] A. G. D. G. B. A. D. P. a. R.-C. T. Andrei-Daniel TUDOSI, «Design and Implementation of an Automated Dynamic Rule System for Distributed Firewalls,» *Advances in Electrical and Computer Engineering,* p. 10, 2023.

Sesión Transferencia II

SCORPION Cybertech: *spin-off* dedicada a la formación en ciberseguridad con un Cyber Range

José A. Ruipérez-Valiente
SCORPION Cybertech
Universidad de Murcia
jruiperez@scorpioncybertech.com

Félix Gómez Mármol
SCORPION Cybertech
Universidad de Murcia
felixgm@scorpioncybertech.com

Félix J. García Clemente
SCORPION Cybertech
Universidad de Murcia
fgarcia@scorpioncybertech.com

***Resumen*—SCORPION Cybertech es una *spin-off* de la Universidad de Murcia dedicada al entrenamiento en ciberseguridad mediante un *Cyber Range* de nueva generación. Esta plataforma integra tres características claves: 1) una herramienta de autor avanzada para la creación de ciberejercicios de diversos tipos, 2) una gestión y monitorización integral del ciclo de vida del ciberejercicio, y 3) un módulo de análisis del desempeño del estudiante mediante la explotación de datos de telemetría y biometría. Todo esto, mediante una arquitectura orientada a microservicios con herramientas y tecnologías de última generación. La solución ha sido diseñada para dar respuesta a desafíos actuales en la formación de profesionales en ciberseguridad, tales como la escasez de personal cualificado, la falta de personalización en el aprendizaje práctico y la baja motivación en entornos formativos tradicionales. En este artículo se presenta el recorrido de transferencia tecnológica que ha dado lugar a la constitución de la empresa, así como las funcionalidades clave del producto, los servicios ofrecidos y su potencial impacto social y educativo.**

***Index Terms*—Cyber Range, entrenamiento en ciberseguridad, gamificación, analítica del aprendizaje, transferencia tecnológica.**

Tipo de contribución: spin-off*s constituidas*

I. Introducción

El acelerado proceso de digitalización de la sociedad ha conllevado un aumento significativo en la superficie de exposición a ciberataques, afectando tanto a infraestructuras críticas como a ciudadanos y pequeñas organizaciones [1], [2]. Esta realidad ha puesto de manifiesto la necesidad urgente de formar profesionales capacitados en ciberseguridad, capaces de operar bajo presión y adaptarse a entornos complejos y dinámicos [3]. La escasez de profesionales especializados en este ámbito ha sido ampliamente reconocida a nivel europeo e internacional [4], generando una fuerte demanda de programas formativos que incluyan entrenamiento práctico basado en escenarios realistas [5].

En este contexto, los *Cyber Ranges* se han convertido en una herramienta esencial para el entrenamiento seguro y realista de competencias en ciberseguridad. Estas plataformas permiten simular redes, sistemas y ataques en entornos virtuales controlados, fomentando el aprendizaje activo y experiencial [6]. No obstante, muchas soluciones existentes presentan limitaciones importantes, como la escasa adaptabilidad de los escenarios, la ausencia de estrategias de gamificación bien diseñadas y la falta de mecanismos de analítica del aprendizaje que permitan personalizar la formación [7], [8], [9].

Además, el diseño instruccional de los ciberejercicios debe contemplar no solo habilidades técnicas, sino también capacidades transversales como la gestión del estrés, la toma de decisiones bajo presión, el trabajo en equipo o capacidades [10]. En este sentido, la incorporación de tecnologías educativas avanzadas, como analíticas multimodales o algoritmos adaptativos, puede mejorar significativamente la experiencia de aprendizaje [10], [11].

En respuesta a estos retos, presentamos SCORPION Cybertech, una *spin-off* de la Universidad de Murcia orientada a la transferencia de tecnología en el ámbito del entrenamiento en ciberseguridad. Su producto principal es un *Cyber Range* innovador que incorpora generación automatizada de escenarios, gamificación adaptativa basada en analíticas del rendimiento, y evaluación de competencias blandas mediante sensores biométricos. Esta solución surge a partir de diversos proyectos de investigación aplicada y representa un ejemplo de cómo el conocimiento generado en el entorno universitario puede transferirse con impacto social, educativo y tecnológico.

El presente artículo describe el recorrido de transferencia tecnológica que ha dado lugar a la constitución de SCORPION Cybertech, destacando las funcionalidades diferenciadoras de su Cyber Range y los servicios que ofrece para distintos perfiles de clientes.

II. SCORPION Cybertech

SCORPION Cybertechnologies S.L. (https://scorpioncybertech.com), o directamente *SCORPION Cybertech*, es una *spin-off* de la Universidad de Murcia que se constituyó en el primer trimestre de 2025 y cuyos socios fundadores son:

- Félix Gómez Mármol, Director Ejecutivo
- Félix J. García Clemente, Director Técnico
- José A. Ruipérez Valiente, Director de I+D

La Figura 1 muestra el logotipo definido para SCORPION Cybertech.

Figura 1. Logotipo de SCORPION Cybertechnologies S.L.

II-A. Antecedentes

En 2020 la Dirección General de Armamento y Material (DGAM) del Ministerio de Defensa de España seleccionó el proyecto “COBRA – Cibermaniobras adaptativas y personalizables de simulación hiperrealista de APTs y entrenamiento

en ciberdefensa usando gamificación"[1] entre las propuestas financiables de su programa COINCIDENTE para desarrollar un prototipo de Cyber Range para el Mando Conjunto del Ciberespacio (MCCE).

Por su parte, la Fundación Séneca de la Región de Murcia concedió en 2022 el proyecto "Smart and efficient Cyber Range-as-a-Service for on-demand Trainings"[2] dentro de su convocatoria de Pruebas de Concepto, lo que sirvió para aumentar el TRL (Technology Readiness Level) de la solución de Cyber Range.

Finalmente, en 2023 la DGAM financió nuevamente el proyecto "COBRA 2.0"[3] como una evolución del anterior que permitiera elevar su TRL hasta ponerlo en producción en el MCCE.

Todos los proyectos anteriormente mencionados han sido desarrollados por la Universidad de Murcia, y más particularmente el Laboratorio de Ciberseguridad y Ciencia de Datos (CyberDataLab[4]) y en particular han estado involucrados los tres socios de la *spin-off*. Esto nos ha permitido adquirir el conocimiento necesario para poder constituir la *spin-off* de cara a proveer estos servicios a distintos actores del sistema social, tanto públicos como privados.

II-B. SCORPION Cyber Range

SCORPION Cyber Range presenta un conjunto distintivo de características funcionales avanzadas, cuidadosamente diseñadas para maximizar el potencial formativo y operativo del entorno. Estas funcionalidades se articulan en torno a tres roles diferenciados de usuario, cada uno con responsabilidades específicas dentro de la plataforma:

- **Administrador**: Posee control completo del sistema, encargándose de la gestión integral del entorno del Cyber Range, con visibilidad total y acceso sin restricciones a todas las funciones disponibles.
- **Instructor**: Responsable directo de la creación, gestión y supervisión de los ciberejercicios. Define las topologías, configura escenarios, establece retos específicos, asigna ejercicios a los alumnos y realiza seguimiento activo durante su ejecución.
- **Estudiante**: Es el usuario destinatario final de los ciberejercicios, realizando las actividades asignadas en un entorno virtual controlado, monitorizado y aislado.

Las características funcionales destacadas de SCORPION Cyber Range son:

1. *Herramienta Avanzada de Autoría para la Creación de Ciberejercicios*
   SCORPION dispone de una herramienta especializada que permite al instructor definir escenarios complejos mediante un editor visual intuitivo, como se muestra en la Figura 2. Esta herramienta facilita la creación precisa y detallada de la topología del ciberejercicio, ajustándose al contexto deseado por el instructor. Por ejemplo, en escenarios orientados a la Industria Inteligente (Smart Industry), pueden incorporarse fácilmente nodos específicos como sistemas SCADA, PLC o sensores IoT. En escenarios tipo Smart Office, se pueden integrar elementos como estaciones de trabajo, routers o firewalls.
   Además, esta herramienta posibilita la introducción de entidades no jugadoras (NPCs, Non-Player Characters) que ejecutan comportamientos benignos automatizados, generando tráfico realista y simulando actividad típica de usuarios legítimos. Asimismo, se permite la incorporación de Amenazas Persistentes Avanzadas (APTs), configuradas mediante timelines específicos para desencadenar acciones maliciosas controladas durante el ejercicio.
   Los retos diseñados siguen una estructura basada en Capture The Flag (CTF). Cada reto se resuelve al capturar una "flag"definida por el instructor. La herramienta permite definir no solo esta bandera, sino también incluir pistas graduadas con penalizaciones asociadas, asignar puntuaciones variables según la dificultad y clasificar los retos en categorías técnicas específicas (p.ej., Crypto, Web, Reverse Engineering). El instructor controla además la visibilidad y accesibilidad de nodos específicos en función del reto en curso.
   Importante destacar que las topologías y retos se definen y almacenan independientemente de los ciberejercicios específicos, facilitando su reutilización en múltiples escenarios futuros. Al crear un nuevo ciberejercicio, el instructor selecciona una topología previamente definida, elige los retos específicos y determina el tipo y formato del ejercicio, pudiendo elegir entre múltiples modalidades: ofensiva (red-team), defensiva (blue-team), combinada (purple-team), individual, colaborativa o adaptativa.
2. *Gestión y Monitorización Integral del Ciclo de Vida del Ciberejercicio*
   Una vez diseñado un ciberejercicio, la plataforma ofrece opciones avanzadas para su despliegue inmediato o para su programación en fechas y horarios específicos. Cada despliegue se realiza de forma aislada, asegurando un entorno independiente para cada alumno o equipo participante, lo que garantiza la integridad y privacidad del proceso formativo.
   Durante la ejecución del ciberejercicio, los alumnos acceden mediante escritorios virtualizados, disponiendo únicamente de acceso a los nodos definidos explícitamente por el instructor. La plataforma monitorea en tiempo real el desempeño del alumno, registrando cada acción relevante, tales como la resolución de retos, solicitudes de pistas o el tiempo empleado.
   El instructor cuenta con herramientas de visualización gráfica avanzada que proporcionan información actualizada del progreso individual y colectivo. Adicionalmente, puede interactuar activamente con los entornos virtuales desplegados, accediendo directamente a cualquier nodo y ejecutando acciones en tiempo real para ajustar o supervisar dinámicamente el ejercicio.
3. *Análisis y Explotación Detallada del Desempeño del Alumno*
   Al concluir un ciberejercicio, SCORPION genera au-

[1] https://portalinvestigacion.um.es/proyectos/586581/detalle
[2] https://portalinvestigacion.um.es/proyectos/587060/detalle
[3] https://portalinvestigacion.um.es/proyectos/701879/detalle
[4] https://cyberdatalab.um.es/

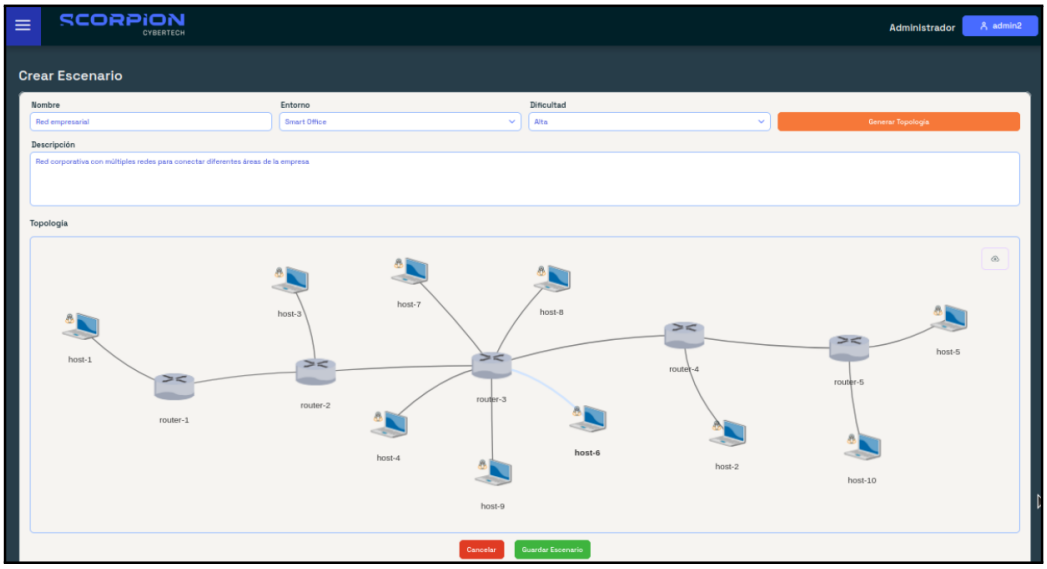


Figura 2. Pantalla para la creación de un escenario en SCORPION Cyber Range.

tomáticamente registros exhaustivos que capturan el rendimiento individualizado del alumno, incluyendo métricas como la puntuación total obtenida, tiempos dedicados por reto, pistas utilizadas y acciones críticas realizadas.

Estos registros permiten generar estadísticas detalladas que facilitan la evaluación integral del aprendizaje. La plataforma incorpora mecanismos de gamificación mediante la asignación de medallas virtuales como reconocimiento de logros específicos definidos por el instructor, como superar ciertos umbrales de puntuación, resolver retos específicos o participar activamente en ejercicios colaborativos.

Una funcionalidad diferenciadora de SCORPION es su capacidad de análisis segmentado por categoría técnica de reto. Este análisis segmentado permite a la plataforma ofrecer ciberejercicios adaptativos, ajustando automáticamente su dificultad y contenido en función del desempeño previo registrado, mejorando así el aprendizaje personalizado y progresivo del alumno.

En cuanto a su arquitectura, SCORPION Cyber Range se fundamenta en un diseño orientado a microservicios, lo que le otorga importantes ventajas técnicas. Esta aproximación arquitectónica facilita una escalabilidad horizontal eficiente, permitiendo añadir o modificar funcionalidades de forma modular sin afectar al funcionamiento del sistema en su conjunto. Además, contribuye a la resiliencia del sistema, ya que cada servicio puede ser desplegado, actualizado o recuperado de manera independiente, y mejora la mantenibilidad del código al separar responsabilidades de forma clara.

El despliegue de los ciberejercicios se gestiona a través de un orquestador que automatiza la creación de entornos virtualizados sobre máquinas virtuales o contenedores. Las máquinas virtuales proporcionan un alto grado de aislamiento y compatibilidad con sistemas operativos completos, siendo útiles en ejercicios donde se requiere simular entornos realistas. Sin embargo, su elevado consumo de recursos y mayor tiempo de inicialización pueden ser limitantes cuando se trata de escalar a muchos usuarios simultáneamente. En contraste, el uso de contenedores permite despliegues más ágiles y ligeros, con un uso más eficiente del hardware disponible, lo que resulta clave cuando se ejecutan escenarios complejos con múltiples nodos para un gran número de alumnos.

Otro aspecto arquitectónico destacado es la integración con sistemas de gestión del aprendizaje mediante el estándar LTI (Learning Tools Interoperability). Este estándar, promovido por IMS Global, permite que SCORPION actúe como proveedor LTI y se integre fácilmente con plataformas LMS (Learning Management System) como Moodle u OpenEDX. Esta integración posibilita una experiencia de aprendizaje unificada, en la que los ciberejercicios pueden formar parte natural de itinerarios educativos más amplios, facilitando la trazabilidad del progreso del estudiante y la combinación con otros recursos.

Finalmente, la arquitectura de SCORPION incorpora un plano de control que proporciona capacidades avanzadas de monitorización en tiempo real. Esta capa controla tanto la infraestructura física como los despliegues virtuales específicos de cada alumno, permitiendo detectar y responder rápidamente a fallos o desviaciones en el comportamiento del sistema. Esta capacidad no solo mejora la fiabilidad y seguridad de la plataforma, sino que también facilita una supervisión activa por parte del administrador, quien puede intervenir en tiempo real durante el uso de la plataforma, garantizando así una experiencia educativa controlada y estable.

II-C. Servicios disponibles y clientes potenciales

Los principales servicios ofrecidos desde SCORPION Cybertechnologies pueden resumirse en:

1. **Cursos prediseñados de ciberseguridad**: Incluye paquetes de cursos prediseñados para capacitar a equipos en distintos aspectos de la ciberseguridad. Incluye material audiovisual y escenarios prácticos ya construidos que permiten entrenar habilidades clave sin necesidad de adaptar el contenido. Por ejemplo, podrían estar centrados en hacking ético o en la preparación de cierta

cerficación. Es la opción ideal para quienes buscan una formación rápida, efectiva y lista para usar.

2. **Formación personalizada con gemelo digital**: Ofrece un entrenamiento totalmente adaptado a la realidad de tu empresa. A partir de la creación de un gemelo digital de tu infraestructura, desarrollamos ciberejercicios de ataque y defensa personalizados que permiten entrenar a tus equipos en un entorno seguro pero fiel a su realidad operativa. También incluye contenidos formativos específicos para tus necesidades y objetivos.
3. **Cyber Range as a Service (CRaaS)**: Permite a la empresa utilizar directamente nuestra plataforma para diseñar y ejecutar sus propios ciberejercicios. Proporcionamos acceso completo al entorno simulado junto con soporte técnico, mientras tú gestionas los contenidos y escenarios según tus propios objetivos. Es la solución perfecta para organizaciones con capacidad interna que buscan autonomía y flexibilidad.

Por su parte, los tipos de clientes a los que principalmente está dirigida la solución de SCORPION Cyber Range son:

- **Pequeñas y medianas empresas (PYMEs)**. Nuestra solución puede formar a los equipos de TI de empresas que no se pueden permitir grandes inversiones
- **Fuerzas y cuerpos de seguridad**. Para ayudar a identificar y combatir a cibercriminales y ciberdelincuentes.
- **Instituciones académicas**. Para formar a las futuras generaciones de expertos de ciberseguridad en Universidades, centros de FP, institutos e incluso colegios.
- **Administraciones públicas**. a nivel local, autonómico o nacional, como ayuntamientos, consejerías, ministerios, etc.

II-D. Direcciones futuras

En SCORPION Cybertech se plantea mantener una inversión importante en I+D así como la participación en proyectos de innovación que permita mantener nuestro Cyber Range en la vanguardia tecnológica y el estado del arte. Tenemos objetivos de investigación y desarrollo de distintos tipos, pero podemos destacar algunos de ellos. Por ejemplo, queremos ampliar el tipo de ciberejercicio disponibles, incluyendo nuevos tipos como el de supervivencia y el de humano-contra-humano. Además, también planteamos el desarrollo de ciberejercicios focalizados en nuevas temáticas, como aquellos centrados en operaciones de influencia (o desinformación) dentro del dominio cognitivo, así como los centrados en los planos operativo, táctico y estratégico de operaciones en el espectro cibernético. Como parte del desarrollo de los ciberejercicios, también planeamos la integración de un Large Lenguage Model (LLM) que pueda facilitar al instructor la tarea de generar desde 0 ciberejercicios, sirviendo de apoyo y fuente de ideas, además de facilitar ciertas tareas repetitivas.

De forma adicional, también cabe destacar que planeamos el desarrollo de un planificador novel, que permita crear una secuencia temporal con ramificaciones en formato de árbol con interacciones entre los estudiantes, APTs, NPCs y otros eventos dentro de la historia que contextualiza el ciberejercicio. Esto permitirá el desarrollo de cibermaniobras cada vez más cercanas a las que se desarrollan en entornos reales.

Por último, también planeamos expandir el procesado de datos multimodal, incorporando ciberejercicios colaborativos donde se recoja información de los canales de comunicación de los miembros del equipo con el objetivo de hacer un análisis detallado de la misma, por ejemplo, para evaluar al líder del equipo durante la ejecución del ciberejercicio.

III. Conclusiones

SCORPION Cybertech representa un caso exitoso de transferencia tecnológica desde la universidad al tejido productivo, aportando una solución avanzada para el entrenamiento en ciberseguridad. Gracias a su arquitectura flexible, su enfoque pedagógico basado en gamificación y analítica del aprendizaje, y su capacidad para generar escenarios adaptativos y realistas, SCORPION Cyber Range se posiciona como una herramienta clave para mejorar la capacitación de profesionales en múltiples sectores. La consolidación de esta *spin-off* refuerza el papel de la investigación aplicada como motor de innovación y compromiso con la seguridad digital de la sociedad.

Agradecimientos

Nos gustaría agradecer a todas las personas que han estado involucradas en el desarrollo de los proyectos en la temática Cyber Range que han dado soporte al proceso de adquisición del conocimiento hasta la constitución de esta *spin-off*.

Referencias

[1] M. Karjalainen and T. Kokkonen, "Requirements for a comprehensive cyber range: A survey of existing platforms," *IEEE Access*, vol. 8, pp. 81 700–81 715, 2020.

[2] T. Rouquette and S. Jajodia, "User traffic generation in cyber ranges: Enhancing realism for cyber defence training," *Journal of Cybersecurity*, vol. 8, no. 1, 2022.

[3] Y. Tzitzikas and N. Papadakis, "An overview of cyber ranges: Concepts, applications and research opportunities," *Computers & Security*, vol. 78, pp. 136–157, 2018.

[4] ENISA, "European cybersecurity skills framework," https://www.enisa.europa.eu/publications/european-cybersecurity-skills-framework, 2020.

[5] G. Hatzivasilis, O. Soultatos, and I. Askoxylakis, "Threat-arrest: A cyber range platform for training against advanced cyber threats," *Future Internet*, vol. 13, no. 5, p. 125, 2021.

[6] R. Beuran, C. D. Pham, I. Chisalita, and H. Simo, "Kypo cyber range: Design and use cases," in *Proceedings of the 2018 International Conference on Cyber Situational Awareness*, 2018.

[7] M. S. Lee and S. H. Yoon, "Icstasy: A cyber range for military-oriented training with automated agents," *Sensors*, vol. 21, no. 12, p. 4052, 2021.

[8] M. Smyrlis, G. Hatzivasilis, and S. Ioannidis, "Cyra: Cyber range assurance platform with model-based training," in *Proc. of the 10th Int. Conference on Modern Circuits and Systems Technologies (MOCAST)*, 2021.

[9] Z. Schreuders and et al., "Secgen: A framework for generating cyber security training scenarios," in *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, 2017, pp. 104–109.

[10] M. Albaladejo-González, P. Nespoli, F. Gómez Mármol, and J. A. Ruipérez-Valiente, "A multimodal and adaptive gamified system to improve cybersecurity competence training," *Cluster Computing*, 2025, in press.

[11] P. Nespoli, M. Albaladejo-González, J. A. P. Valera, J. A. Ruipérez-Valiente, J. Garcia-Alfaro, and F. G. Mármol, "Scorpion cyber range: Fully customizable cyberexercises, gamification and learning analytics to train cybersecurity competencies," *arXiv preprint arXiv:2401.12594*, 2024.

Detección de desinformación en redes sociales: aplicación y validación en un entorno real

Brais Pérez Vázquez
Gradiant
Vigo, España
bperez@gradiant.org

Alberto Fontán Santamaría
Gradiant
Vigo, España
afontan@gradiant.org

Jose Fuentes Rodríguez
Gradiant
Vigo, España
jfuentes@gradiant.org

***Resumen*—El crecimiento del contenido en redes sociales, acompañado por la expansión de la generación y difusión de desinformación, se ha convertido en una preocupación cada vez mayor en la sociedad. Esto supone un desafío fundamental para mantener un ecosistema digital saludable. No obstante, su detección es una tarea compleja, ya que verificar la veracidad de un gran volumen de publicaciones es un proceso costoso, exigente y, en muchos casos, sujeto a la interpretación del analista. En este trabajo, se exploran dos enfoques para abordar este problema: la detección de bots y la identificación de tendencias no orgánicas mediante Inteligencia Artificial. La primera estrategia se centra en identificar cuentas creando contenido de manera automatizada, mientras que la segunda analiza patrones anómalos en la difusión de contenido dentro de las tendencias de las redes sociales. Estas herramientas forman parte del proyecto OSBORNE lo que permite su validación por parte de los clientes finales con datos extraídos de un entorno real.**

***Index Terms*—Tendencias, Detección de Bots, Desinformación, Redes Sociales, Deep Learning.**

Tipo de contribución: Transferencia

I. Introducción

En la actualidad, la difusión de desinformación en las redes sociales se ha convertido en un problema de gran trascendencia, con consecuencias profundas que repercuten en la opinión pública. La lógica algorítmica de estas plataformas tiende a priorizar contenidos sensacionalistas, lo que potencia la propagación de información falsa y refuerza narrativas engañosas. Además, el anonimato y las campañas coordinadas exacerban la situación, permitiendo que la desinformación alcance audiencias masivas y adquiera una legitimidad artificial.

Ante este panorama, resulta imprescindible implementar nuevos mecanismos y herramientas que faciliten la detección de información falsa. El desarrollo de herramientas para la detección de desinformación tiene un gran potencial en la protección del ecosistema digital, beneficiando a los usuarios de redes sociales, medios de comunicación y las diferentes fuentes de información. En este proceso, la Inteligencia Artificial (IA) y el análisis avanzado de datos juegan un papel clave, facilitando la identificación de patrones anómalos en la propagación de contenidos y la detección de campañas de manipulación. El uso de herramientas basadas en IA evita procesos manuales en tiempo real, un proceso complejo, costoso y que puede verse afectado por la subjetividad del analista. Al integrarse en plataformas digitales y redes sociales, estas herramientas permiten distinguir el origen de los focos de desinformación, permitiendo una moderación más precisa y reduciendo la amplificación de información engañosa.

En este trabajo, se presenta una herramienta basada en IA para la detección de cuentas con una actividad automatizada, a traves de modelos de grafos, y la detección de tendencias con picos de actividad no orgánicos, a través de *forecasting* de series temporales. Los modelos de IA descritos son un claro ejemplo de transferencia tecnológica ya que, se integran en la herramienta desarrollada por la empresa CORRECTA en el proyecto OSBORNE, financiado mediante los fondos de la Unión Europea NextGeneration-EU a través del Instituto Nacional de Ciberseguridad (INCIBE)[1], una entidad dependiente del Ministerio para la Transformación Digital y de la Función Pública y la Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales. El objetivo del proyecto es la mejora del proceso de análisis e investigación de redes sociales por parte de organismos gubernamentales. Todos los datos utilizados en la validación de los modelos implementados provienen por tanto de entornos reales, recolectados bajo la supervisión y la aceptación de los usuarios finales de la herramienta. Este enfoque garantiza que la investigación académica se traduzca en soluciones prácticas de alto valor, demostrando la aplicabilidad y el impacto de la tecnología en escenarios operativos concretos.

II. Caso de Uso

Las estrategias usadas en este trabajo para la detección de la desinformación en las redes sociales se basan en la detección de *bots* o cuentas automatizadas, y la detección de patrones anómalos en las tendencias, que son los tópicos discutidos en diferentes periodos de tiempo. El desarrollo de la herramienta se centra en la red social *X*, anteriormente conocida como Twitter, pero es extrapolable de manera directa a plataformas similares como *Threads*, *Bluesky* o *Mastodon*.

Las cuentas automatizadas en redes sociales, comúnmente llamadas *bots*, suelen asociarse a la generación de grandes volúmenes de contenido sobre temas específicos con el fin de obtener mayor visibilidad. Si somos capaces de identificarlas, podremos reconocer las principales fuentes de creación de contenido en determinadas temáticas. En la actualidad, el auge de la IA y los modelos de lenguaje hace que el comportamiento de estas cuentas se parezca cada vez más al de los usuarios humanos. Por ello, para detectarlas, ya no basta con analizar únicamente su actividad, sino también observar cómo se relacionan con otros usuarios de la red y de qué manera estos últimos interactúan con sus publicaciones.

La detección de cuentas automatizadas en redes sociales ha sido un tema de estudio ampliamente abordado en los últimos años, especialmente en la plataforma *X*, donde su

[1] https://www.incibe.es/

presencia es significativamente mayor en comparación con otras redes sociales. Como resultado, se han desarrollado numerosos modelos con el objetivo de detectar *bots*. Estos modelos constituyen una base fundamental para nuestro trabajo, y, además, algunas de las bases de datos utilizadas en su entrenamiento son de acceso público. Entre las bases de datos disponibles, TwiBot-22 [1] se destaca como la más reciente y completa, además de ser la que cuenta con el mayor número de modelos entrenados sobre ella. Esta base de datos contiene más de un millón de usuarios etiquetados y más de ochenta millones de publicaciones, lo que la convierte en un recurso de referencia para la investigación en este ámbito.

Por otro lado, en redes sociales como *X*, una tendencia está compuesta por las publicaciones de usuarios en determinados instantes de tiempo, las cuales mencionan entidades que se encuentran dentro del ámbito de la tendencia mediante palabras clave, cadenas de texto o *hashtags*. Además, el concepto de tendencia está ligado al de popularidad, pues no se suele considerar que un tema es tendencia hasta que se discute en un número de publicaciones significativo y durante un periodo de tiempo sostenido, lo cual también conlleva que la tendencia tenga un mayor alcance y una mayor capacidad de diseminar información. Así pues, a partir de recopilar las publicaciones de una tendencia y en un rango de tiempo, se puede analizar dicha tendencia en base al comportamiento de diferentes características de las publicaciones a lo largo del marco de tiempo, como usuarios participantes, número de publicaciones por unidad de tiempo, interacciones con las publicaciones como *reposts*, *likes* y comentarios, análisis del contenido de las publicaciones, entre muchas otras.

En materia de desinformación, esto es relevante cuando algún *topic* con objetivo de desinformar es una tendencia con un considerable alcance y popularidad, en cuyo caso se podrían tomar medidas para evaluar y mitigar el impacto. De esta forma, se pueden definir dos tipos de tendencias:

- **Orgánicas**: tendencias más o menos virales pero con un comportamiento esperado, donde su popularidad vaya creciendo y decreciendo de determinada forma dependiendo de las características de la tendencia.
- **No orgánicas**: tendencias cuyo comportamiento no es el esperado, con picos injustificados que no responden a ningún patrón de cómo los humanos suelen interaccionar con tendencias de las mismas características.

En la literatura, se han propuesto múltiples contribuciones, tanto directas como indirectas, para el análisis de tendencias en redes sociales, en particular en *X*. Estos trabajos abordan diversas temáticas, que van desde la identificación de características determinantes en la popularidad de una publicación ([2], [3], [4]) y el uso de modelos de *Machine Learning* para la predicción de series temporales [5], hasta la detección y clasificación de tendencias orgánicas y alteradas [6]. Si bien estas investigaciones ofrecen una visión general del campo y proporcionan un punto de partida sólido, sus enfoques no coinciden de manera exacta con el modelado que aquí se plantea, lo que impide establecer comparaciones directas con trabajos previos. Además, la propia naturaleza dinámica de los datos, aunada a la variabilidad de las plataformas y los contextos temporales y sociales, dificulta la consolidación de un estado del arte unificado.

III. Metodología

En esta sección se describe la metodología seguida para la implementación de los modelos de IA de detección de *bots* y detección de tendencias no orgánicas.

III-A. Detección de Bots

Con el objetivo de detectar la actividad de cuentas automatizadas, utilizaremos redes neuronales de grafos. Estas redes se basan en representar la dinámica de interacción en la red social mediante un grafo, donde los nodos corresponden a los usuarios involucrados en una tendencia y las aristas capturan las interacciones entre ellos.

Concretamente, se emplean *Graph Attention Networks* (GAT) [7], un tipo de redes de grafos que utiliza un mecanismo de atención para asignar diferentes pesos a las conexiones entre usuarios en función de su relevancia. Este modelo ajusta de manera iterativa la importancia de las relaciones durante el proceso de aprendizaje, lo que mejora su capacidad para capturar patrones distintivos. Además, procesará el grafo para asignar a cada usuario una puntuación que indica la probabilidad de que sea un *bot*. A continuación, se detalla la información contenida en los grafos:

1. ***Nodos***: usuarios que participan en la tendencia. Se calculan 60 características sobre los usuarios y sus publicaciones, como la distribución del número de mayúsculas por publicación, la distribución de las horas de las publicaciones, la proporción de seguidores/seguidos o la antigüedad de la cuenta. Además se obtiene un *embedding* de la descripción de la cuenta y un *embedding* de las últimas publicaciones.
2. ***Aristas***: cada una de las interacciones entre dos usuarios (*repost*, menciones, citados y respuestas). Dentro de cada arista se almacena la información del número de relaciones de cada tipo entre cada par de usuarios y el número de *hashtags* iguales usados por ambos.

III-B. Detección de tendencias no orgánicas

Para abordar la detección de tendencias no orgánicas, se plantea el problema como un problema de detección de anomalías en series temporales. En este enfoque, las características de las publicaciones realizadas en momentos específicos se utilizan como variables predictoras, mientras que el número total de publicaciones se considera la variable respuesta. A continuación, se entrena un modelo de IA que, a partir de dichas variables predictoras, infiere el valor de la variable de respuesta y lo compara con el dato real, lo que permite determinar si existe alguna anomalía.

En este caso se ha implementado una Red Neuronal Recurrente (RNN) [8] con celdas Long Short Term Memory (LSTM) [9] que procesa secuencias de datos de una longitud determinada y produce un único valor de salida. Dicha secuencia de entrada es una secuencia de ventanas de tiempo, donde cada una contiene las características agregadas de las publicaciones realizadas en ese marco de tiempo. Por ejemplo, si se agregan publicaciones por ventanas de dos minutos, y la tendencia comienza el 01/01/2025 a las 12:00, las ventanas contendrán las características de las publicaciones realizadas entre 12:00-12:02, 12:02-12:04, 12:04-12:06, y así consecutivamente, como se puede ver en la *Figura* 1.

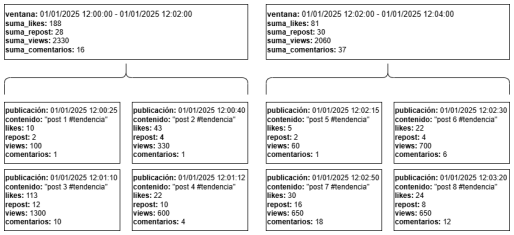


Figura 1. Muestra de ventanas de publicaciones

Así pues, el modelo toma *N* ventanas como secuencia de entrada, y se optimiza para inferir el número de publicaciones de la ventana *N+1*, pasando por todas las ventanas de la tendencia. De esta forma, se puede entrenar el modelo sobre tendencias orgánicas para aprender los patrones que identifican tendencias con un comportamiento no alterado, de tal manera que cuando procese una secuencia de una tendencia alterada, obtenga un error más alto entre el valor inferido para la ventana de salida y el valor real. A partir del error obtenido en cada ventana de la tendencia, y de un umbral de detección previamente definido, es posible detectar las ventanas que se corresponden con una tendencia no orgánica.

IV. Impacto y Resultados

Una vez descrita de manera general las estrategias empleadas y la importancia de las mismas para el desarrollo de la herramienta , en esta sección se muestran los resultados obtenidos por ambos modelos y la importancia e impacto de los mismos.

IV-A. Detección de bots

Como se mencionó en la *Sección* II, la base de datos utilizada para la detección de *bots* en *X* es TwiBot-22. Dado que nuestro desarrollo está orientado a una posible transferencia de tecnología, hemos considerado las recientes modificaciones en la política de privacidad de *X*. Por ello, a diferencia de otros enfoques en la literatura, hemos empleado únicamente la información actualmente accesible en *X*, evitando datos que ya no están disponibles, como las listas de seguidores/seguidos o los usuarios que interactúan con publicaciones mediante *likes*. Cabe destacar que los resultados expuestos en la tabla de los modelos del estado del arte, están evaluados sobre la información no disponible actualmente, para una evaluación más justa deberían estar probados sobre el mismo conjunto de datos.

Tabla I
Resultados obtenidos en comparación a los mejores modelos de la literatura

Modelo	*F1-score*	*Accuracy*	*Precision*	*Recall*
BotRGCN [10]	0,5750	0,7966	0,7481	0,4680
Random Forest [11]	0,5870	0,7640	0,786	0,4680
GAT (0.5)	0,5757	0,7197	0,5193	0,6457
GAT (0.6)	0,5677	0,7399	0,5560	0,5798

La *Tabla* I muestra los resultados obtenidos por el modelo propuesto (GAT), usando distintos thresholds, 0.5 y 0.6, para determinar si es bot o no, respecto a las dos mejores aproximaciones disponibles en el estado del arte con el mismo *dataset*, existiendo otros 25 algoritmos que obtienen peores resultados. Podemos observar como las métricas obtenidas por el modelo son muy similares, centrandonos en el F1-score, la puntuación obtenida es casi la misma, pero destaca la capacidad de nuestro modelo para detectar un mayor número de bots (un 65 % y 57 % del total frente un 46 %), debido a esto, genera también un mayor número de falsos positivos (el 51 % y 55 % de los usuarios detectados como bots realmente lo son, frente al 75 %). No se tratan de unas métricas ideales, sin embargo, se podrían ajustar, como podemos observar, modificando el threshold.

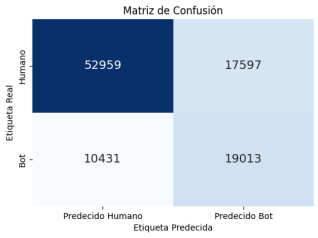


Figura 2. Matriz de Confusión

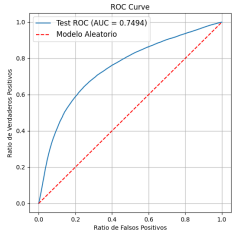


Figura 3. Curva ROC

La matriz de confusión (véase en la *Figura* 2) muestra que la herramienta identifica eficazmente cuentas con actividad automatizada, lo que sugiere que prioriza la detección de cuentas con alto riesgo de ser *bots*. Esto puede generar algunos falsos positivos, pero dado que el objetivo es facilitar la investigación de estas cuentas, este margen de error es aceptable. La curva *ROC* (véase en la *Figura* 3) refleja la relación entre la tasa de verdaderos y falsos positivos en diferentes umbrales. El *AUC-ROC* de 0.741 indica un rendimiento prometedor en la detección de *bots*, ya que muestra una capacidad adecuada de discriminación entre cuentas automatizadas y humanas.

IV-B. Detección de tendencias no orgánicas

Como comentábamos anteriormente en la *Sección* II, en cuanto a la detección de anomalías en tendencias de redes sociales, no se cuenta con un estado del arte previo sólido que marque claramente la línea de investigación a seguir y permita la comparación rigurosa entre diferentes enfoques, por lo que la evaluación de nuestra aproximación se lleva a cabo de la siguiente manera: se toman los datos de publicaciones de *X* de tres tendencias orgánicas y de una tendencia alterada. Estas tendencias se procesan como indicamos anteriormente, agregando las características de las publicaciones por ventanas de tiempo según su fecha de publicación y formando *K* secuencias de *N* ventanas. Posteriormente, se entrena la RNN con las secuencias de ventanas obtenidas de dos de las tendencias orgánicas. Para validar la efectividad del modelo, se realiza la inferencia sobre la tendencia orgánica restante y la tendencia alterada.

Tabla II
Resultados del modelo sobre las tendencias disponibles.

Tendencia	*Mean Absolute Error*
Tendencias orgánicas de entrenamiento	2,01
Tendencia orgánica de prueba	2,48
Tendencia alterada de prueba	5,31

La *Tabla* II muestra el error medio absoluto obtenido en cada uno de estos casos. El modelo es capaz de inferir el

número de publicaciones de una ventana (*N+1*) a partir de la secuencia de *N* ventanas previas, con un error bajo en el caso de la tendencia orgánica de prueba, próximo al error obtenido en entrenamiento, y con un error mucho mayor en el caso de la tendencia alterada. No obstante, el *Mean Absolute Error* no es la única medida relevante para evaluar la calidad de este modelo, ya que se trata de un promedio de los errores obtenidos en todas las ventanas inferidas, y lo normal en una tendencia, aunque sea alterada, es que no todas las ventanas de tiempo sean anómalas y exista una gran cantidad de ellas que corresponda a comportamientos orgánicos. Por ejemplo, si una red de *bots* se coordina durante una ventana de tiempo en concreto para alterar una tendencia, esta experimentará un pico de publicaciones anómalo en dicha ventana y el error en la misma será mucho más alto, pero la media se verá reducida por el resto de ventanas.

Por tanto, el otro indicador a tener en cuenta es el número de ventanas anómalas identificadas cuando el error entre el número de publicaciones inferido y el real supera el umbral. Las ventanas anómalas aparecen señaladas en rojo en las Figuras 4 y 5.

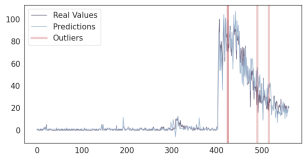


Figura 4. Tendencia Orgánica

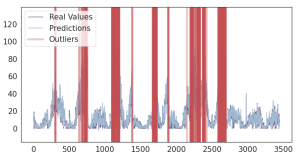


Figura 5. Tendencia No Orgánica

Es así que asumiendo las tendencias de prueba como *Ground Thruth*[2] de que son realmente orgánicas o alteradas en determinados puntos, se puede evaluar si el modelo está identificando correctamente si una ventana tiene un comportamiento orgánico o no. En particular, se observa que se detectan muy pocas anomalías en la tendencia considerada orgánica, mientras que en los picos de publicaciones que se sabe corresponden a ventanas alteradas se identifican muchas más, lo que respalda el comportamiento esperado del modelo.

V. Conclusiones

Los resultados obtenidos confirman la viabilidad de desarrollar una herramienta basada en diversas estrategias para la detección de desinformación en redes sociales. Además de mejorar el análisis de contenido digital, esta herramienta es útil para organismos gubernamentales, permitiendo evaluar la veracidad de tendencias y mitigar la propagación de información engañosa. El campo de estudio en el que se enmarca esta investigación es altamente complejo, lo que implica que las herramientas desarrolladas requieren mejoras continuas para alcanzar una mayor precisión en sus resultados. Sin embargo, los avances logrados hasta ahora trazan un camino prometedor para la investigación y la transferencia en este ámbito, pudiendo seguir mejorando los resultados de los modelos para obtener herramientas más confiables.

El enfoque adoptado en el desarrollo de ambas estrategias de los modelos es innovador y diferencial obteniendo unos resultados no vistos en el estado del arte para el caso de uso de detección de *bots*, e implementando un enfoque innovador en la detección de tendencias. Sin embargo, su relevancia no se limita únicamente a los avances en la investigación, sino que radica en su potencial aplicación y transferencia a la sociedad. Con esta herramienta, es posible realizar un análisis casi inmediato de las tendencias emergentes, facilitando la detección de aquellas que podrían estar influenciadas por una actividad organizada y permitiendo detectar los momentos en los que existe una actividad inusual. Además, permite analizar el comportamiento de los usuarios involucrados, detectando aquellos con una puntuación de actividad automatizada elevada. Aunque este estudio se centró en la red social *X*, los resultados pueden extrapolarse a otras plataformas donde los usuarios interactúan de forma similar (Bluesky o Mastodon). Un desafío clave para futuras investigaciones es la adaptación de estas tecnologías a plataformas con estructuras diferentes, donde la definición de tendencias no son tan directas como en *X*.

[2]Todas las tendencias fueron seleccionadas y proporcionadas por cliente por presentar características y evidencias claras de responder a un comportamiento orgánico o alterado.

Agradecimientos

Esta investigación es parte del proyeto OSBORNE (CPP002-R18), financiado mediante los fondos de la Unión Europea NextGeneration-EU en el marco del Plan de Recuperación, Transformación y Resiliencia (PRTR), a través del Instituto Nacional de Ciberseguridad (INCIBE), entidad dependiente del Ministerio para la Transformación Digital. El proyecto está englobado en la Iniciativa Estratégica de Compra Pública de Innovación (IECPI), un programa impulsado por INCIBE en el marco del PRTR, cuyo objetivo es fomentar la innovación y el desarrollo de soluciones avanzadas en ciberseguridad mediante la contratación pública estratégica.

Referencias

[1] S. Feng et al., "TwiBot-22: Towards Graph-Based Twitter Bot Detection", en *arXiv*, eprint 2206.04564, 2023. [En línea]. Disponible en: https://arxiv.org/abs/2206.04564.

[2] Das, A., Roy, M., Dutta, S., Ghosh, S., & Das, A. (2015). Predicting Trends in the Twitter Social Network: A Machine Learning Approach. pp. 570-581.

[3] Balaji, T. K., Chandra Sekhara Rao Annavarapu, & Annushree Bablani. (2021). Machine learning algorithms for social media analysis: A survey. Computer Science Review, 40.

[4] Silva, M., Dominguez, M., & Celayes, P. (2018). Analyzing the Retweeting Behavior of Influencers to Predict Popular Tweets, with and without Considering their Content.

[5] Han, Z., Zhao, J., Leung, H., Ma, K. F., & Wang, W. (2021). A Review of Deep Learning Models for Time Series Prediction. IEEE Sensors Journal, 21(6), 7833-7848. doi: 10.1109/JSEN.2019.2923982.

[6] Assenmacher, D., Clever, L., Pohl, J. S., Trautmann, H., & Grimme, C. (2020). A Two-Phase Framework for Detecting Manipulation Campaigns in Social Media. In G. Meiselwitz (Ed.), Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis. HCII 2020. Lecture Notes in Computer Science, vol. 12194. Springer, Cham. doi: https:/10.1007/978-3-030-49570-114.

[7] P. Veličković et al., "Graph Attention Networks", en *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018. [En línea]. Disponible en: https://arxiv.org/abs/1710.10903.

[8] M. I. Jordan. Serial order: A parallel distributed processing approach. Technical Report ICS Report 8604, Institute for Cognitive Science, University of California, San Diego, 1986

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term memory. Neural computation, 9(8):1735–1780, 1997.

[10] S. Feng, "BotRGCN: Twitter bot detection with relational graph convolutional networks", en *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min. (ASONAM)*, 2021.

[11] M. Kantepe y M. Ganiz, "Preprocessing framework for Twitter bot detection", en *Proc. Int. Conf. Comput. Sci. Eng. (UBMK)*, pp. 630-634, 2017. doi:10.1109/UBMK.2017.8093483.

EUCINF: EUropean Cyber and INFormation warfare toolbox

Alejandro Buitrago López*, Teresa García de Alcaraz Ruiz*, Mario Fernández Tárraga*,
Alejandro D. Cayuela Tudela*, Felipe Sánchez González*, Javier Pastor-Galindo†,
José A. Ruipérez-Valiente*, Gregorio Martínez Pérez*
*Department of Information and Communications Engineering, University of Murcia, 30100, Murcia, Spain
{alejandro.buitragol, teresa.garciaa, mario.fernandezt, alejandrodavid.cayuelat, felipe.sanchezg, jruiperez, gregorio}@um.es
†Computer Systems Engineering Department, Universidad Politecnica de Madrid, Spain
javier.pastor.galindo@upm.es

***Abstract*—In the digital era, influence operations have become increasingly prevalent, solidifying Cyber and Information Warfare (CIW) as a critical strategic domain in modern conflicts, with the European Union (EU) and its Member States increasingly targeted by coordinated disinformation campaigns and influence operations. The EUropean Cyber and INFormation warfare toolbox (EUCINF) presents a distinctive European industrial approach by designing an AI-powered repository of configurable applications alongside an experimental platform to tackle the challenges of digital information operations. Through developing, prototyping, and demonstrating advanced CIW capabilities, EUCINF aims to establish a comprehensive toolbox that integrates interoperable components, a dedicated testbed, and a repository for AI-driven applications, reinforcing Europe's strategic cyber and information warfare capabilities.**

***Index Terms*—Cyber defense, Artificial Intelligence, Disinformation, Information Warfare**

Type of contribution: *Research in development*

I. Introduction

In an increasingly digitalized and globalized world, information has emerged as the second most significant instrument of power, surpassed only by the national economy [1]. Beyond its civilian impact, information plays a crucial role in military strategy, as the rise of influence operations has positioned cyber and information warfare among the most strategic domains of modern conflicts [2], [3].

One of the primary methods employed in these influence operations is disinformation, which has been defined as false information designed to mislead [4]. Far from being an isolated phenomenon, disinformation is a deliberate and systematic strategy designed to serve multiple objectives. By distorting and manipulating the truth, such campaigns aim to undermine societal trust, destabilize institutions, and advance specific geopolitical interests [5].

Also, while the use of manipulated information in conflict is not a new phenomenon, what is unprecedented today is the emergence of what some experts have termed **"cyber-enabled information warfare."** This modern form of conflict is characterized by the strategic and intentional use of information to target adversaries. The goal is to foster confusion, spread false narratives, or manipulate adversaries' decision-making processes [6], [7].

The EU and its member states have increasingly become prime targets of coordinated influence operations and disinformation campaigns [8]. In response to these growing threats, numerous cybersecurity initiatives have been proposed in recent years. Among them, frameworks such as DISARM have been developed to adapt cybersecurity models for better detection and prevention of such incidents, strengthening resilience against the evolving challenges of information warfare [9], [10].

Building on these efforts, EUCINF has been established to provide dedicated support tailored to counter the threats posed by cyber-enabled information warfare.

EUCINF brings together 22 partners, from 12 countries, in a joint collaboration over a 36-month period (01/12/2023 – 01/12/2026), through which EUCINF will design, prototype, and demonstrate advanced CIW capabilities through a comprehensive toolbox that integrates interoperable components, a specialized testbed, and a repository of AI-driven, configurable applications. Doing so will serve as a reference platform for innovation and operational effectiveness in cyber and information warfare.

Bringing a distinctive European industrial approach, EUCINF focuses on developing an AI-powered repository of configurable applications alongside an experimental platform to tackle the complexities of digital information operations. By addressing three operational CIW scenarios aligned with 9 Ministry of Defense (MoD) requirements, legal frameworks, and societal expectations, EUCINF establishes the technical foundation for a European CIW toolbox. This ensures strategic, operational, and tactical effectiveness in both civil and military cyber and information warfare contexts.

The rest of the paper is structured as follows. In Section II, related projects are presented. Section III describes the EUCINF framework. Section IV presents the discussion. Finally, we end the paper with conclusions in Section V.

II. State of the art

Detecting and documenting disinformation campaigns at scale is essential for understanding the evolving tactics, techniques, and procedures (TTPs) used in information manipulation operations [9]. In this context, several EU-funded projects have emerged recently to address disinformation and counter its effects.

A. Related projects

In the civil domain, disinformation has been actively tackled by the EU institutions since 2015 through many initiatives,

such as the creation of the East Stratcom Task Force and its flagship project EUvsDisinfo [11], a database of articles and media which the organization considers as providing false, distorted, or partial information.

One group of projects primarily focused on detecting, analyzing, and countering Foreign Information Manipulation and Interference (FIMI) through structured methodologies and frameworks. By way of example, the ADAC.io (Attribution, Data Analysis, Countermeasures, and Interoperability) [12] project aims to protect democracy by trying to alleviate the effects of FIMI on society with the use of guided and structured approaches such as the COE Attribution Framework or the DISARM Framework.

This last framework is also used by the ATHENA [13] project, which focuses on studying more than 30 FIMI case studies and their used TTPs, resulting in informative knowledge graphs and disinformation analysis. In this vein, the SAUFEX (Secure Automated Unified Framework for EXchange) [14] project addresses the need for harmonization of the approach towards FIMI by proposing the creation of a FIMI dataspace.

Another group of projects examines information suppression by authoritarian states as a form of FIMI, aiming to provide policy recommendations and defensive strategies. For example, the ARM [15] project aims to study the increase in power of autocracies with the use of certain TTPs by involving a defender community. Similarly, the DE-CONSPIRATOR [16] project aims at a more precise response by methodologically evaluating the most successful FIMI cases. In contrast, the RESONANT [17] project tries to also address this objective by focusing on identifying TTPs.

B. Innovative character of EUCINF

The projects above explore different approaches to combating disinformation and influence operations from a purely civilian perspective. While some proposals focus on analyzing the environments where disinformation occurs, others concentrate on categorizing disinformation campaigns using TTPs or enhancing existing frameworks. These approaches have resulted in public reports and datasets that, although valuable, are only available after incidents have taken place.

EUCINF, however, introduces a civil-military framework that addresses disinformation from a CIW perspective. This project provides a comprehensive military workflow solution, covering everything from information collection and exploitation—transforming data into cyber intelligence to developing countermeasures and decision support in counter-disinformation operations.

This framework has developed specialized tools for each phase, ensuring high interoperability. These tools can work seamlessly with one another and with external Cyber Threat Intelligence (CTI) sharing platforms. In addition to incorporating the functionalities of previous projects, EUCINF enhances them with advanced mission analysis and countermeasure development capabilities, distinguishing itself as a unique civil-military initiative.

III. EUCINF FRAMEWORK

The overall framework architecture is depicted in Fig. 1, illustrating the structured approach EUCINF follows in its operations. EUCINF will work under a generic Cyber & Information Warfare operational workflow encompassing **data collection**, **analysis**, **processing**, and **capitalization**.

This workflow ensures that information is systematically gathered, transformed into actionable intelligence, and utilized within the broader project objectives. This approach will benefit the development, deployment, and use of the EUCINF ToolBox, enhancing its capabilities to support data-driven decision-making and strategic responses.

A. Objectives

The EUCINF ToolBox aims to significantly enhance European **capabilities in cognitive warfare** (Scenarios) through the development of a **shared library** (Store) of innovative, configurable **software products** (Set of components). A dedicated Interoperability framework, its associated Testbed, and Common Functions to **secure and control the use of the components** provided will ease their integration and reuse into existing CIW systems or demonstration environments. Fig. 1 depicts the overall workflow.

Each software component within the EUCINF Toolbox is designed to perform an **autonomous task**, carrying out a specific function defined by its inputs and outputs. These components are built to be interoperable, ensuring seamless integration and replacement within a **modular and adaptable** EUCINF workflow. This design allows for incorporating third-party components, meaning external elements not originally included in the Toolbox, as long as they adhere to standardized interfaces and operational behaviors.

The development of these components follows the principles, specifications, methods, and guidelines established by the EUCINF Interoperability Framework, ensuring consistency and compatibility across the system. Their interoperability will be rigorously tested and validated within the EUCINF Testbed using **standardized testing procedures on designated test datasets**. These datasets will be provided by the owners of the tested components, ensuring alignment with best practices in AI development and compliance with **ethical and legal frameworks** governing responsible AI deployment.

The EUCINF Framework will be developed based on a coherent set of relevant open standards, ensuring consistency and interoperability across its components. The EUCINF Testbed will allow for evaluating and validating these standards, their effectiveness, and how they function together in an integrated environment.

Finally, the EUCINF Toolbox will leverage cross-cutting technical capabilities to support demonstration activities and provide **common services**, including a **Store** with a registry. This registry will serve as a structured repository where components and datasets are cataloged using **standardized descriptions and interfaces**, allowing project partners to access shared content and contribute their own. To facilitate seamless integration, the registry will offer its services through a standardized interface, enabling efficient component configuration and the exchange of tool specifications and characteristics among users.

B. Innovation and technological development

The EUCINF Toolbox is designed to be a reference repository of AI-based configurable applications as a testbed for

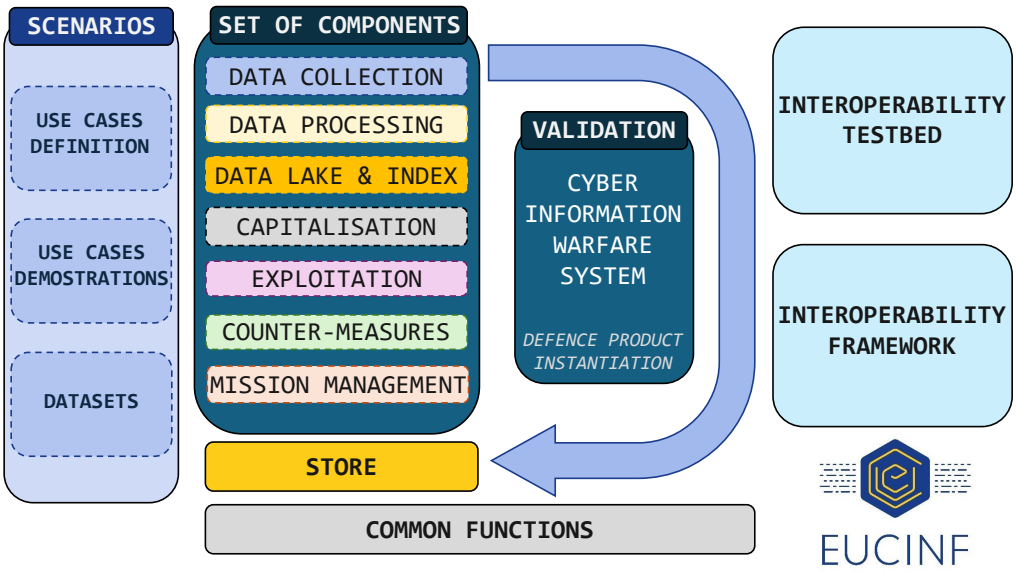


Fig. 1. Global view of the EUCINF solution logic.

creating and experimenting with artificial intelligence techniques ad hoc for Cyber Security and Information Warfare. Its primary objective is to address the particular challenges of these contexts, particularly in disinformation tracking applications, by exploiting state-of-the-art AI-powered solutions. By integrating various AI tools and processes, the Toolbox will enhance the detection, analysis, and mitigation of cyber threats and adversarial information operations, enabling a stronger and more adaptive security stance.

To this end, the Toolbox is designed to address several operational dimensions, integrating and orchestrating various technologies with capabilities to process multi-source information in CIW operations. Among its core competencies, it will possess information processing mechanisms to provide early notice of adversary information operations and activities. It will also possess Cyber Threat Intelligence tools and procedures to effectively counter blended malicious cyber and information operations accepted as integral components of contemporary hybrid warfare approaches. Additionally, the Toolbox will provide situational awareness of adversary information operations, informing operational decisions with real-time insight into emerging threats. One of the key functions of the Toolbox will be to establish bridges to other operations, particularly cyber operations, to foster a synchronized and holistic response to emerging threats.

Beyond its functional competencies, the Toolbox will include rules, methodologies, and tools to facilitate the development, integration, orchestration, and sharing of Toolbox assets, including data, modules, and applications. The framework will ensure stakeholders can deploy and customize AI solutions effectively while guaranteeing interoperability and adaptability in dynamic operational environments.

In parallel with its basic functionality, the development of the Toolbox will be supported by detailed research on standardization, interoperability guidelines, and privacy protection. This will ensure that the Toolbox is aligned with existing regulatory frameworks and best practices, making it simple to integrate into existing cyber defense systems while adhering to ethical and legal requirements. By addressing these technical and strategic dimensions, the EUCINF Toolbox is intended to be a cornerstone in AI-driven cyber defense and information security, fostering innovation while building resilience against emerging threats in the cyber battlefield.

IV. Discussion

Although the EUCINF focuses on enhancing European capabilities in cyber and cognitive warfare defense, it must still comply with European legal and ethical standards (European Defence Fund Regulation EU 2021/697). To ensure this, it undergoes audits that assess its proper implementation and verify that its use remains strictly for defense, avoiding any unethical or illegal purposes.

One of the most critical aspects of compliance is adherence to the Compliance with the General Data Protection Regulation (GDPR) (Regulation EU 2016/679), particularly when collecting and processing data from open sources. Special attention must be given to personal and sensitive data to ensure ethical and legal adherence. To address these concerns, clear protocols have been established for data collection, processing, retention, and deletion. Additionally, mechanisms such as anonymization and pseudonymization play a crucial role in safeguarding individual privacy while maintaining data utility.

Furthermore, the development of AI solutions must align with best practices and comply with EU legal and ethical frameworks, such as the AI Act (Regulation EU 2024/1689). This ensures that AI systems operate within a structured and accountable environment. A key aspect of this compliance is adherence to the principles of transparency and traceability, allowing for clear documentation and understanding of the decisions made by AI applications.

Also, while AI systems are designed to support decision-making processes, ultimate accountability must always rest with human authority. AI should assist in analysis and rec-

ommendations, but responsible individuals must review and approve final decisions. This human oversight ensures ethical responsibility, reduces risks of automated errors, and fosters public trust in AI-driven solutions.

V. Conclusions

As the Cyber and Information domains increasingly play a decisive role in anticipating and managing both current and future conflicts, the CIW Toolbox concept developed by the EUCINF Consortium is designed to effectively counter the threats emerging from evolving cyber and hybrid warfare capabilities, such as disinformation and deepfakes.

This Toolbox will offer defense end-users a comprehensive library of software-configurable components that can be seamlessly integrated into CIF systems, addressing the need for holistic resilience measures, including threat detection and appropriate countermeasures.

Through its development, the EUCINF project aims to significantly enhance the performance of CIF systems across EU Member States, strengthening their overall defense effectiveness and fostering greater cooperation in cyber defense in alignment with the EU Capability Development Plan Priorities.

Acknowledgements

Funded by the European Union

EUCINF project is co-funded by European Union under grant agreement N°101121418.

Disclaimer

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] A. Crowther, "Reorganizing for information competition," in *ICCWS 2018 13th International Conference on Cyber Warfare and Security*. Academic Conferences and publishing limited, 2018, p. 131.

[2] F. I. Roumpies and A. P. Kakarountas, "Cybersecurity and democracy: A review," in *2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*. IEEE, 2023, pp. 1–8.

[3] J. Pastor-Galindo, P. Nespoli, J. A. Ruipérez-Valiente, and D. Camacho, "Influence operations in social networks," 2025. [Online]. Available: https://arxiv.org/abs/2502.11827

[4] A. B. López, J. Pastor-Galindo, and J. A. Ruipérez-Valiente, "Frameworks, modeling and simulations of misinformation and disinformation: A systematic literature review," 2024. [Online]. Available: https://arxiv.org/abs/2406.09343

[5] W. L. Bennett and S. Livingston, "The disinformation order: Disruptive communication and the decline of democratic institutions," *European journal of communication*, vol. 33, no. 2, pp. 122–139, 2018.

[6] H. Lin and J. Kerr, "On cyber-enabled information/influence warfare and manipulation," *Center for International Security and Cooperation, Stanford, SAD*, pp. 4–22, 2017.

[7] A. B. López, J. Pastor-Galindo, and J. A. Ruipérez-Valiente, "Llm-assisted topic modeling for hate speech characterization," Oct. 2024. [Online]. Available: http://dx.doi.org/10.22541/au.172966882.21215291/v1

[8] M. Theocharidou, "Enisa's threat landscape 2024," 2021.

[9] E. Kapantai, A. Christopoulou, C. Berberidis, and V. Peristeras, "A systematic literature review on disinformation: Toward a unified taxonomical framework," *New media & society*, vol. 23, no. 5, pp. 1301–1326, 2021.

[10] F. S. González, J. Pastor-Galindo, and J. A. Ruipérez-Valiente, "Toward interoperable representation and sharing of disinformation incidents in cyber threat intelligence," 2025. [Online]. Available: https://arxiv.org/abs/2502.20997

[11] "EUvsDisinfo — Detecting, analysing, and raising awareness about disinformation - EUvsDisinfo — euvsdisinfo.eu," https://euvsdisinfo.eu, [Accessed 30-01-2025].

[12] "ADAC.IO: Attribution, Data Analysis, Countermeasures and Interoperability — adacio.eu," https://adacio.eu, [Accessed 27-02-2025].

[13] "Project Athena — project-athena.eu," https://project-athena.eu, [Accessed 27-02-2025].

[14] "Saufex - Secure Automated Unified Framework for Exchange — saufex.eu," https://saufex.eu, [Accessed 27-02-2025].

[15] "Arm project — www.arm-project.eu," https://www.arm-project.eu, [Accessed 27-02-2025].

[16] "Home - DE-CONSPIRATOR — deconspirator.eu," https://deconspirator.eu, [Accessed 27-02-2025].

[17] "Home - Resonant Project — resonantproject.eu," https://resonantproject.eu, [Accessed 27-02-2025].

Generación de títulos en español para la alerta temprana de vulnerabilidades mediante Inteligencia Artificial Generativa

Javier Román Martínez*, David Triana Robles‡, Mouhcine El Oualidi Charchmi§, Ines Salamanca Estévez†, Noemí DeCastro-García¶
RIASC, Universidad de León
Campus de Vegazana s/n, 24071 León, España
* jromm@unileon.es, ‡ dtrir@unileon.es, § mouac@unileon.es, † isale@unileon.es, ¶ ncasg@unileon.es

***Resumen*—La gestión de la alerta temprana de vulnerabilidades desempeña un papel clave en el ámbito de la ciberseguridad, ya que permite minimizar el tiempo de exposición a posibles amenazas. Este estudio analiza el impacto de incluir sistemas de inteligencia artificial en la generación automatizada de títulos para alertas tempranas gestionadas por el Instituto Nacional de Ciberseguridad. Se evalúan diferentes herramientas de inteligencia artificial generativa, comparando la precisión, la coherencia semántica y la eficiencia de los títulos generados. A través del uso de modelos de representación vectorial de palabras y análisis estadístico, se identifican patrones en la estructura de los títulos y se determina qué herramienta ofrece los mejores resultados.**

***Index Terms*—Inteligencia Artificial Generativa, Ciberseguridad, Vulnerabilidades, Generación automática de contenido**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

La proliferación de vulnerabilidades en sistemas informáticos ha convertido la gestión de la alerta temprana en un elemento esencial para la ciberseguridad. Identificar y comunicar vulnerabilidades de manera eficiente permite a las organizaciones aplicar medidas preventivas rápidamente, minimizando el impacto de posibles ataques. Dado que los títulos son la primera referencia en la gestión de vulnerabilidades, una correcta estandarización de los mismos mejora la detección, análisis y respuesta temprana a incidentes de seguridad. En este contexto, la generación precisa y eficiente de títulos para reportes de vulnerabilidades es un factor clave para la correcta identificación y comprensión de los riesgos asociados.

El Instituto Nacional de Ciberseguridad (INCIBE) es responsable de la gestión de la alerta temprana en español de vulnerabilidades, un proceso que actualmente se lleva a cabo de manera manual. Este procedimiento comienza cuando el NIST (National Institute of Standards and Technology [1]) publica nuevas vulnerabilidades en la base de datos NVD (National Vulnerability Database [2]), proporcionando una descripción detallada en inglés. INCIBE accede a esta información y la incorpora a su propio *feed* de vulnerabilidades, el cual posteriormente es publicado en su web [3]. De forma paralela a esta integración, en la plataforma del NIST, INCIBE lleva a cabo la traducción manual de la descripción de la vulnerabilidad y la generación de un título adecuado en español. Una vez que esta fase ha concluido, la información traducida y revisada se actualiza en el *feed* interno y, finalmente, en la web de INCIBE, garantizando que los usuarios tengan acceso a una versión en español de las alertas de vulnerabilidades.

Este proceso, aunque funcional, resulta laborioso, lo que puede provocar demoras y variaciones en la calidad de los títulos generados. La aplicación de Inteligencia Artificial Generativa (IAG) en este contexto ofrece una oportunidad para optimizar la generación de títulos, garantizando mayor coherencia, reduciendo los tiempos de procesamiento y mejorando la eficiencia en la difusión de alertas de seguridad. Para que la aplicación de la IAG sea efectiva, es fundamental que el problema a abordar presente patrones de comportamiento que puedan ser identificados y aprendidos por los modelos [4]. En estos casos, los modelos pueden generalizar contenido a partir de ejemplos previos, optimizando su capacidad de respuesta y reduciendo la incertidumbre en sus predicciones. Al observar que los títulos siguen estructuras recurrentes, surge la idea de explorar la automatización del proceso mediante IAG.

En este artículo se presenta una evaluación preliminar de la eficiencia que presentan distintas herramientas de IAG en la generación de títulos de la alerta temprana en español de vulnerabilidades. Un aspecto fundamental en el proceso de generación de títulos es la traducción de las descripciones de las vulnerabilidades en las que se basan. Estas traducciones son esenciales para la construcción de los títulos. En estudios previos [5], se ha analizado la generación automática de traducciones y su impacto en la precisión y coherencia del contenido generado. De los resultados obtenidos, se ha extraído la selección de IAGs que se han analizado en este estudio: ChatGPT [6], Copilot [7] y ALIA [8]. La primera parte de la investigación se ha centrado en identificar patrones en la creación de títulos que permitieran obtener el *prompt* óptimo que sirva como *input* a cada IAG. Seguidamente, se ha pasado a generar los títulos de una muestra de vulnerabilidades para, finalmente, determinar la herramienta más adecuada para la tarea mediante análisis estadístico y de embebimiento de palabras, estableciendo criterios objetivos que faciliten su implementación en la gestión de alertas de vulnerabilidades.

Este artículo se organiza de la siguiente manera: la Sección II describe el diseño del experimento. En la Sección III, se presentan los hallazgos experimentales. Finalmente, se incluyen las conclusiones de este trabajo de investigación.

II. Sección experimental

En esta sección se describe el diseño experimental, incluyendo las preguntas de investigación, los datos utilizados, la selección de IAGs y la metodología para identificar patrones, generar títulos y evaluar su calidad mediante similitud

semántica.

II-A. Preguntas de investigación

El objetivo general de este trabajo es determinar el efecto de incluir IAG en el proceso de generación de títulos de la alerta temprana de vulnerabilidades que realiza INCIBE. Las preguntas de investigación que se quieren resolver en este trabajo se enumeran a continuación:

- PI1 ¿Existen patrones en los títulos de las descripciones de las vulnerabilidades? ¿Qué patrón es el más adecuado/-frecuente/útil?
- PI2 ¿Qué IAG aporta los títulos de mejor calidad?
- PI3 ¿Qué IAG es más eficiente en la tarea de generación de títulos para las descripciones de la alerta temprana de vulnerabilidades?

II-B. Conjuntos de datos

Se utilizó un *dataset* proporcionado por INCIBE [9], el cual contiene un total de 205282 vulnerabilidades, cada una con su respectiva descripción en inglés, su traducción al español y su título. Este conjunto de datos sirvió como base para la identificación de patrones en los títulos de las vulnerabilidades y el análisis de similitud semántica.

Debido a las limitaciones de las IAGs, que impedían la lectura de archivos demasiado grandes, se estudiaron específicamente las primeras 2269 vulnerabilidades con título para la identificación de patrones de forma automática.

Por otro lado, para llevar a cabo las pruebas de generación de títulos, se seleccionó una muestra compuesta por las primeras 22 vulnerabilidades con título registradas en la base de datos. Esta selección se realizó con el objetivo de garantizar que, dentro de las restricciones propias de las IAG utilizadas, el tiempo de procesamiento no resultara excesivo. Esto se debió a que, tanto por las limitaciones de las versiones gratuitas de las herramientas comparadas como por las del hardware disponible, la generación de respuestas por parte de cada IAG implicaba un consumo considerable de recursos y tiempo.

II-C. IAGs

En este estudio se emplean diversas IAGs con el objetivo de generar los títulos en español de las alertas tempranas de vulnerabilidades. La generación de títulos depende de forma inherente a la descripción de las vulnerabilidades. Por ese motivo, la selección se ha basado en otras investigaciones relacionadas [5]. Así, las traducciones han sido generadas por los sistemas de traducción automático e IAGs que se enumeran a continuación:

- ChatGPT [6] (Versión GPT-4o mini): Modelo de lenguaje avanzado desarrollado por OpenAI, capaz de realizar traducciones y generar texto en múltiples idiomas con alto nivel de fluidez y contexto.
- Bing Translator [10]: Servicio de traducción basado en IA de Microsoft, integrado en diversas plataformas y aplicaciones.
- DeepL [11]: Herramienta de traducción que emplea redes neuronales avanzadas para ofrecer traducciones de alta calidad con especial énfasis en matices lingüísticos.
- Google Translate [12]: Traductor automático de Google, ampliamente utilizado y mejorado mediante aprendizaje automático y redes neuronales.
- Copilot For Security (CFS) [13]: Versión especializada de Copilot enfocada en la seguridad informática, con capacidades avanzadas de análisis y generación de texto técnico.
- Copilot [7]: Asistente de IA desarrollado por Microsoft basado en modelos de OpenAI, diseñado para asistir en la programación y otras tareas relacionadas con el desarrollo de software.
- CFS1: CFS enriquecido con la base de datos de referencia [9]: Variante de CFS que trabaja con la base de datos utilizada como referencia, incluyendo las muestras analizadas, para mejorar la precisión en tareas específicas.
- CFS2: CFS enriquecido con la base de datos de referencia [9] sin las descripciones de la muestra.
- Salamandra [14]: Herramienta basada en IA utilizada en tareas de generación y traducción de contenido.

Excepto en CFS, CFS1 y CFS2, se han utilizado versiones gratuitas de todas las herramientas. Así, y basándonos en los resultados obtenidos en [5], las IAG seleccionadas para la generación de títulos han sido ChatGPT, Copilot y Salamandra.

Cabe destacar que también se ha llevado a cabo la experimentación partiendo de la descripción general en inglés.

II-D. Metodología

El flujo de trabajo incluyó:

1. Preprocesamiento de los datos de la muestra de INCIBE [9]: El procesamiento de los datos se llevó a cabo mediante un *script* en *Python* que extrajo todos los títulos existentes, eliminó duplicados y los guardó en un archivo de texto plano. Posteriormente, se realizó un análisis manual para identificar patrones en los títulos determinando inicialmente la existencia de seis estructuras distintas. Para realizar este análisis se tuvo en cuenta que los sistemas de alerta temprana suelen basarse en la recolección y análisis de informes de vulnerabilidades provenientes de diversas fuentes, como bases de datos de seguridad y reportes de investigadores (ej. National Vulnerability Database, NVD [2]; Common Vulnerabilities and Exposures, CVE [15]). Para facilitar su clasificación y priorización, los títulos de las vulnerabilidades deben ser claros y estructurados, incluyendo información clave como: Componente afectado (software, dispositivo, módulo), tipo de vulnerabilidad (XSS, SQL Injection, RCE, etc.), versión afectada (si aplica) y contexto del impacto (cómo y dónde se manifiesta la vulnerabilidad).
   Debido al gran número de títulos y el riesgo de error humano, se utilizó *ChatGPT 3.5* y *Copilot*, para ampliar el análisis y validar los grupos identificados manualmente. Esto permitió identificar patrones adicionales, refinando los grupos hasta obtener un total de nueve estructuras.
2. Generación de títulos mediante IAG: se solicitó a las IAGs generar un *prompt* óptimo para la creación de nuevos títulos basados en la descripción de una vulnerabilidad con una estructura que siguiera el patrón seleccionado. Con base en este *prompt*, se generaron

los títulos correspondientes de la muestra; es decir, se generaron títulos para 22 traducciones en español de vulnerabilidades. Este proceso se repitió con cada conjunto de traducciones creadas con los diferentes traductores automáticos explicados en la sección anterior y para la descripción original en inglés.

3. Cálculo de similitud semántica usando *word embeddings* [16].
4. Evaluación de los resultados con los títulos originales propuestos por INCIBE. Este enfoque permitió obtener una evaluación cuantitativa sobre la calidad de los títulos generados, asegurando una metodología robusta y reproducible para futuros estudios.

II-E. Análisis

Para evaluar la similitud semántica entre los títulos generados y los títulos originales, se utilizó la técnica de *word embeddings* junto con la similitud del coseno. Se entrenaron varios modelos de *word embeddings*, empleando diferentes corpus disponibles:

- Primer modelo: basado en todas las traducciones del conjunto de datos original de INCIBE [9].
- Segundo modelo: basado en un corpus de traducciones público proporcionado por INCIBE, el cual fue desarrollado en 2019 en colaboración con la National Vulnerability Database (NVD) del NIST con el objetivo de traducir al español las descripciones de vulnerabilidades recogidas en dicha base de datos. Este esfuerzo resultó en la creación de un corpus bilingüe con más de 100000 unidades de traducción, disponible públicamente en el repositorio ELRC-SHARE, y orientado a mejorar las capacidades de análisis y respuesta ante vulnerabilidades en el ámbito hispanohablante [17].
- Tercer modelo: utilizando modelos preentrenados (`wiki.en.bin` y `wiki.es.bin`) disponibles en *FastText* [18] y en el proyecto de investigación *MUSE* [19] desarrollado por Meta.

Para los dos primeros modelos, se creó un diccionario de *palabras ancla* extrayendo palabras comunes entre las traducciones en inglés y español. Este procedimiento se llevó a cabo mediante un *script* en *Python* que analizó y comparó los archivos de texto, identificando palabras coincidentes y registrando su frecuencia de aparición. Para el tercer modelo, se utilizó un diccionario de *palabras ancla* proporcionado por las fuentes de los modelos.

Cada modelo fue alineado mediante la técnica de Procrustes ortogonal [20], permitiendo calcular la similitud semántica entre las frases comparadas. La similitud se evaluó en un rango de [0,1], donde 0 indica la menor aproximación y 1 la mayor.

Una vez aplicados los *word embeddings*, se calcularon estadísticos descriptivos de los resultados obtenidos de toda la muestra.

II-F. Detalles Técnicos

El procesamiento y análisis de datos se realizaron con herramientas de *Python*, incluyendo las librerías:

- *FastText [18]*: Para el entrenamiento de modelos de *word embeddings*.
- *SciPy [21]*: Para el cálculo de la similitud del coseno.
- *NumPy* [22] y *Pandas* [23]: Para la manipulación y análisis de datos.

III. Resultados

III-A. PI1: Búsqueda de patrones en los títulos

Los patrones fueron analizados primero de forma manual, creando nuevas categorías en caso de que un título no encajara en ninguna de las existentes. Para el análisis con *Copilot* y *ChatGPT*, se utilizó el siguiente *prompt*, adjuntando un archivo JSON con los títulos a analizar:

> *Analiza cómo se han creado los títulos de estos textos, estudia si existen patrones.*

Los patrones encontrados, de forma manual, han sido:

1. Nombre de producto/software/dispositivo
   - Patrón: [Nombre del Producto/Software/Dispositivo]
   - Ejemplo: SAP CRM WebClient UI
2. Código de CVE
   - Patrón: [CVE]
   - Ejemplo: CVE-2018-20249
3. Descripción de vulnerabilidad
   - Patrón: vulnerabilidad de [tipo de vulnerabilidad] en [producto]
   - Ejemplo: Vulnerabilidad de inyección SQL en el módulo Sections para PHP-Nuke
4. Descripción de una función/archivo/módulo
   - Patrón: el/la [nombre de la función/archivo/módulo] en [contexto]
   - Ejemplo: el encabezado HTTP en los dispositivos VIVOTEK IP Camera
5. Componente específico en aplicaciones
   - Patrón: el/la [componente] en [aplicación]
   - Ejemplo: el formulario de inicio de sesión en Microweber CMS
6. Configuraciones y parámetros en aplicaciones/dispositivos
   - Patrón: el/los [parámetro(s)] en [aplicación/dispositivo]
   - Ejemplo: el parámetro system_time_timezone en la función SetNTPServerSettings en D-Link COVR 1200,1202,1203
7. Referencias a componentes y módulos en contextos específicos
   - Patrón: el/la [componente/módulo] en [contexto]
   - Ejemplo: el componente Outside In Technology en Oracle Application Server
8. Operaciones y acciones en productos/aplicaciones
   - Patrón: el/la [acción] en [producto/aplicación]
   - Ejemplo: la generación del código HTML relacionado con los emojis en GitLab CE/EE
9. Nombres específicos de funciones y métodos
   - Patrón: la/el [nombre de la función/método] en [contexto]
   - Ejemplo: la función ext4_xattr_check_entries en el kernel de Linux

A pesar de identificar estos patrones, no se ha encontrado un motivo claro para seleccionar un patrón específico u otro al generar el título basado en su descripción. Cabe destacar que a excepción de los 3 primeros grupos, los títulos siguen siempre la misma estructura:

El/La/Los/Las [nombre de elemento/s afectado/s] en [contexto en el que se ven afectados]

Aunque la fórmula "vulnerabilidad de [tipo de vulnerabilidad] en [producto]" proporciona suficiente información para considerarla una estructura de título válida, tiene una representación muy limitada.

Una vez determinado el patrón óptimo para la generación de títulos, se compartió este análisis con *Copilot* y *ChatGPT*, junto con el siguiente *prompt* para que cada uno generara su propia propuesta de *prompt* ideal:

> *Crea un prompt que pueda utilizarse para generar nuevos títulos a partir de nuevos textos.*

Las respuestas obtenidas se presentan en la Tabla I.

IAG	*Prompt* **propuesto**
ChatGPT	Genera un título conciso para un informe de vulnerabilidad basado en la siguiente descripción. El título debe comenzar con el nombre del archivo, software o sistema afectado, seguido de su ubicación o contexto si es relevante. No menciones la vulnerabilidad específica ni detalles técnicos complejos. Usa un formato similar a estos ejemplos: *'archivo_en_sistema'*, *'módulo en software'*, *'software en versión_afectada'*.
Copilot	Crea un título para un informe de vulnerabilidad siguiendo estos pasos: ▪ Identifica el componente o software afectado. ▪ Especifica el tipo de vulnerabilidad (e.g., XSS, SQL Injection, Buffer Overflow). ▪ Incluye la versión del software si es relevante. ▪ Utiliza preposiciones como *"en"* para indicar el contexto y *"para"* para indicar el propósito. ▪ Añade detalles adicionales entre paréntesis si es necesario.

Tabla I
Prompts PROPUESTOS PARA LA GENERACIÓN DE TÍTULOS DE VULNERABILIDADES

Tras la generación de los *prompts* óptimos por parte de *ChatGPT* y *Copilot*, se evaluó manualmente el desempeño de cada modelo en la generación de títulos, considerando criterios de precisión, claridad y alineación con la descripción de la vulnerabilidad. Cada modelo generó cinco títulos por *prompt*, y su adecuación fue analizada cualitativamente según estos criterios.

El *prompt* generado por *Copilot* fue seleccionado como el más adecuado, ya que sus respuestas presentaban una estructura más coherente con las instrucciones dadas. Además, cuando la descripción de la vulnerabilidad estaba en inglés, se especificó en el *prompt* que el título debía generarse en español para evitar generaciones incorrectas.

III-B. PI2: Evaluación de la calidad de los títulos

Los resultados obtenidos de realizar la similitud del coseno de los *word embedding* se pueden ver reflejados en las tablas II, III y IV

Con respecto al modelo 1, Copilot y Salamandra tienen un rendimiento superior, mientras que ChatGPT obtiene puntuaciones más bajas. Sin embargo, con el modelo 2, Salamandra

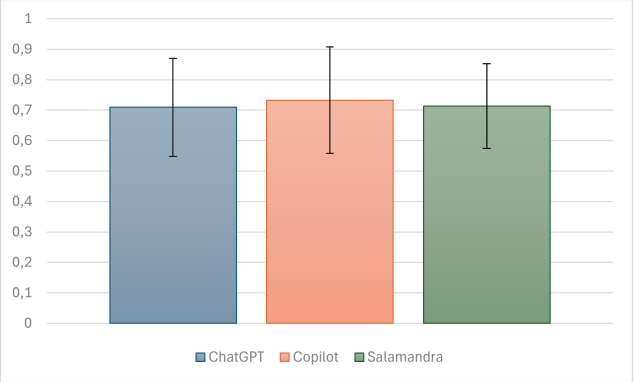


Figura 1. Media de los resultados obtenidos con los tres modelos, comparando los títulos generados a partir de la descripción traducida por INCIBE

lidera, seguido de Copilot, y ChatGPT se mantiene como el menos preciso. Por último, con el modelo 3, Copilot sigue siendo el mejor, con ChatGPT en último lugar. Las diferencias en los resultados pueden deberse a varios factores. Cada modelo de *word embeddings* representa el significado de las palabras de manera diferente, dependiendo de cómo fue entrenado. El modelo 1 está entrenado con vulnerabilidades traducidas a español por INCIBE, por lo que probablemente está más alineado con textos técnicos y de ciberseguridad. Por ese motivo, los títulos generados por modelos más especializados (como Copilot y Salamandra) pueden haberse ajustado mejor al contexto técnico, aportando mejores puntuaciones. Sin embargo, ChatGPT podría haber generado títulos más generales, lo que explicaría su menor rendimiento en este modelo. Por otra parte, el modelo 2 está entrenado con un corpus de ciberseguridad del año 2019, por lo que podría tener una representación semántica más amplia y menos actual. Con este modelo, Salamandra obtiene mejores resultados, lo que podría sugerir que su método de generación de títulos es más adaptable. ChatGPT sigue con puntuaciones más bajas. El modelo 3 está preentrenado con gran cantidad de textos diversos, sin un enfoque específico en ciberseguridad. En este escenario, Copilot sería el mejor generado, lo que indica que es un generador de títulos más robusto en contextos generales. ChatGPT y Salamandra tienen puntuaciones más bajas, lo que sugiere que sus estrategias de generación pueden no estar alineadas con el conocimiento general aprendido por este modelo.

De manera general, aunque los resultados puedan parecer similares (véase la Figura 1), se observan algunas diferencias en la media y la desviación típica obtenidas. No obstante, dado que estas diferencias no parecen significativas, se procede a analizar el coste de implementación en términos de tiempo. Este análisis permite identificar que Copilot presenta un rendimiento más estable y elevado en todos los modelos, lo que refleja su capacidad para extraer términos clave y estructurar los títulos con precisión. Además, se adapta bien tanto a modelos específicos como generales, lo que sugiere una buena consistencia semántica en su algoritmo. Por otro lado, Salamandra muestra un mejor desempeño en modelos específicos, lo que podría indicar que su generación de títulos depende más del contexto técnico. ChatGPT, en cambio, obtie-

Tabla II
RESULTADOS DE LA PRUEBA DE *word embeddings* USANDO EL MODELO ENTRENADO EN ESPAÑOL CON LAS VULNERABILIDADES TRADUCIDAS POR INCIBE.

Traductor	Generador de título	Media	Desv. típica	Mediana	Máximo	Mínimo	Tiempo medio (ms)
INCIBE	ChatGPT	0.7356	0.1552	0.7682	0.9161	0.3325	0.1322
	Copilot	0.7521	0.1709	0.8350	0.9833	0.3492	0.1343
	Salamandra	0.7444	0.1326	0.7796	0.9317	0.3809	0.1837
ChatGPT	ChatGPT	0.7236	0.1596	0.7392	0.9279	0.2871	0.1524
	Copilot	0.7430	0.1697	0.7707	0.9206	0.3018	0.1254
	Salamandra	0.7506	0.1479	0.7745	0.9480	0.3107	0.1474
Bing	ChatGPT	0.7162	0.1605	0.7377	0.9222	0.3001	0.1408
	Copilot	0.7267	0.1610	0.7798	0.9071	0.3172	0.1171
	Salamandra	0.7449	0.1297	0.7773	0.8939	0.3189	0.1534
DeepL	ChatGPT	0.7155	0.1575	0.7353	0.9211	0.3062	0.1323
	Copilot	0.7399	0.1694	0.7743	0.9282	0.3455	0.1205
	Salamandra	0.7502	0.1293	0.7866	0.8916	0.3029	0.1363
Google Translate	ChatGPT	0.7211	0.1537	0.7159	0.9265	0.3222	0.1317
	Copilot	0.7353	0.1668	0.7499	0.9724	0.3139	0.1200
	Salamandra	0.7253	0.1252	0.7500	0.8827	0.3699	0.1392
CFS	ChatGPT	0.7147	0.1596	0.7253	0.9115	0.3222	0.1266
	Copilot	0.7258	0.1668	0.7718	0.9100	0.3139	0.1142
	Salamandra	0.7562	0.1339	0.7821	0.9241	0.4432	0.1414
Copilot	ChatGPT	0.7102	0.1514	0.7000	0.9115	0.3222	0.1304
	Copilot	0.7432	0.1775	0.7990	0.9606	0.3022	0.1264
	Salamandra	0.7376	0.1611	0.7921	0.9106	0.2233	0.1329
CFS1	ChatGPT	0.7261	0.1520	0.7199	0.9273	0.3248	0.1281
	Copilot	0.7529	0.1740	0.8092	0.9760	0.3018	0.1309
	Salamandra	0.7310	0.1154	0.7359	0.8983	0.3557	0.1387
CFS2	ChatGPT	0.7334	0.1630	0.7595	0.9225	0.3008	0.1334
	Copilot	0.7423	0.1720	0.8141	0.9206	0.3018	0.1191
	Salamandra	0.7361	0.1308	0.7700	0.9010	0.4486	0.1509
Salamandra	ChatGPT	0.7720	0.1266	0.7952	0.9325	0.4026	0.1466
	Copilot	0.7864	0.1290	0.8054	0.9657	0.4386	0.1433
	Salamandra	0.7106	0.1322	0.7286	0.8823	0.2840	0.1485
Descripción original en inglés	ChatGPT	0.7899	0.1240	0.8298	0.9346	0.4038	0.1490
	Copilot	0.8107	0.1183	0.8399	0.9502	0.4136	0.1348
	Salamandra	0.7477	0.1148	0.7739	0.9172	0.3927	0.1964

ne el rendimiento más bajo en todas las pruebas, posiblemente debido a que genera títulos más creativos, lo que reduce su alineación semántica con la descripción de referencia.

Por otro lado, resulta evidente que, en general, la generación automatizada de títulos funciona mejor cuando se parte de la descripción original de las vulnerabilidades en inglés, véase Tabla V. Al partir de la traducción a español de la descripciones, se podría estar perdiendo información o alterando la estructura semántica de la oración. Aunque las diferencias no parecen significativas, la descripción original en inglés permite una mejor generación del título. Este resultado podría ser debido a que tanto ChatGPT como Copilot han sido entrenadas con una mayor cantidad de datos en inglés que en español, comprendiendo y procesando mejor los textos en este idioma. Es por ello que la diferencia no es tan acusada cuando utilizamos Salamandra.

Por otra parte, el análisis mediante *word embeddings* podría estar ocasionando sesgos debido a que estos modelos de procesamiento de lenguaje natural están mejor calibrados para detectar relaciones semánticas en inglés que en español.

III-C. PI3: Eficiencia y limitaciones

Salamandra-7b-instruct tiene un tiempo de generación de título por cada descripción medio de 43.33 segundos mientras que Copilot genera las traducciones en un tiempo medio de 0.91 segundos y ChatGPT en 1.42 segundos.

Al realizar las pruebas de generación de títulos con Salamandra-7b-instruct, se ha observado un peor rendimiento. A pesar de que el *prompt* instrucciones detalladas, Salamandra-7b-instruct no genera los títulos siguiendo dichas indicaciones. En numerosos casos, es necesario postprocesar la respuesta, ya que reescribe la descripción de la vulnerabilidad en lugar de crear un título adecuado, o bien produce títulos genéricos que requieren ser completados manualmente. Esto obliga a repetir el proceso de generación para obtener resultados utilizables. Estos problemas no se han observado en ChatGPT o Copilot.

IV. CONCLUSIONES

El uso de una asistente basado en IAG puede mejorar la eficiencia en el proceso de generación automática de títulos en español de la alerta temprana de vulnerabilidades.

Como conclusiones finales, Copilot es el mejor generador de títulos en términos de precisión. Salamandra ofrece resultados sólidos y consistentes y ChatGPT genera títulos menos alineados. Además, en base a los resultados obtenidos, se recomendaría realizar este proceso de manera previa a realizar la traducción a español.

Como trabajo futuro, sería relevante ampliar la muestra de estudio, tanto en cantidad como en el intervalo de fechas considerado, con el fin de comprobar si los ajustes se degradan a medida que se incrementa el volumen de contenido y analizar

Tabla III
RESULTADOS DE LA PRUEBA DE *word embeddings* USANDO EL MODELO ENTRENADO EN ESPAÑOL CON EL CORPUS DE 2019.

Traductor	Generador	Media	Desv. típica	Mediana	Máximo	Mínimo	Tiempo Medio
INCIBE	ChatGPT	0.7405	0.1490	0.7788	0.9135	0.3968	0.1302
	Copilot	0.7566	0.1655	0.8257	0.9779	0.3857	0.1340
	Salamandra	0.7589	0.1198	0.7852	0.9348	0.4778	0.1915
ChatGPT	ChatGPT	0.7339	0.1503	0.7532	0.9177	0.3730	0.1502
	Copilot	0.7513	0.1651	0.7715	0.9119	0.3712	0.1243
	Salamandra	0.7665	0.1305	0.7982	0.9478	0.4083	0.1543
Bing	ChatGPT	0.7221	0.1499	0.7446	0.9201	0.3794	0.1394
	Copilot	0.7349	0.1537	0.7846	0.9115	0.3992	0.1171
	Salamandra	0.7600	0.1166	0.7884	0.8943	0.3948	0.1591
DeepL	ChatGPT	0.7247	0.1510	0.7521	0.9075	0.3815	0.1304
	Copilot	0.7458	0.1627	0.7799	0.9071	0.3825	0.1192
	Salamandra	0.7614	0.1114	0.7756	0.9068	0.4488	0.1422
Google Translate	ChatGPT	0.7282	0.1485	0.7324	0.9152	0.4076	0.1311
	Copilot	0.7428	0.1619	0.7491	0.9693	0.3898	0.1188
	Salamandra	0.7425	0.0999	0.7441	0.9032	0.4672	0.1432
CFS	ChatGPT	0.7221	0.1521	0.7260	0.8928	0.4076	0.1260
	Copilot	0.7334	0.1579	0.7780	0.9081	0.3898	0.1139
	Salamandra	0.7729	0.1173	0.7916	0.9161	0.5119	0.1480
Copilot	ChatGPT	0.7186	0.1474	0.7261	0.8928	0.4076	0.1295
	Copilot	0.7593	0.1708	0.8061	0.9620	0.3747	0.1249
	Salamandra	0.7617	0.1236	0.7982	0.9082	0.4184	0.1367
CFS1	ChatGPT	0.7326	0.1447	0.7249	0.9173	0.4041	0.1266
	Copilot	0.7580	0.1665	0.7876	0.9707	0.3712	0.1290
	Salamandra	0.7455	0.1040	0.7616	0.9047	0.4223	0.1478
CFS2	ChatGPT	0.7440	0.1516	0.7859	0.9152	0.3853	0.1317
	Copilot	0.7448	0.1639	0.7949	0.9044	0.3712	0.1177
	Salamandra	0.7526	0.1071	0.7760	0.8907	0.5280	0.1540
Salamandra	ChatGPT	0.7899	0.1205	0.8215	0.9388	0.4936	0.1449
	Copilot	0.8036	0.1105	0.8149	0.9644	0.5307	0.1421
	Salamandra	0.7340	0.1096	0.7662	0.8864	0.4307	0.1573
Descripción original en inglés	ChatGPT	0.8029	0.1111	0.8303	0.9392	0.4752	0.1536
	Copilot	0.8288	0.1105	0.8546	0.9496	0.4606	0.1407
	Salamandra	0.7637	0.1043	0.7947	0.8972	0.4573	0.2039

cómo se comporta la generación de títulos en vulnerabilidades más recientes.

AGRADECIMIENTOS

Esta investigación resulta del Proyecto Estratégico *Ciencia de datos para un modelo de inteligencia artificial en ciberseguridad* (C073/23), fruto del convenio de colaboración suscrito entre el Instituto Nacional de Ciberseguridad (INCIBE) y la Universidad de León. Esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation).

REFERENCIAS

[1] NIST, "NIST," https://www.nist.gov/, Último acceso: 15 de enero de 2025.

[2] ——, "National vulnerability database," 2005, Último acceso: 21 de enero de 2025. [Online]. Available: https://nvd.nist.gov/

[3] INCIBE, "Vulnerabilidades," https://www.incibe.es/incibe-cert/alerta-temprana/vulnerabilidades, Último acceso: 7 de marzo de 2024.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[5] Javier Román Martínez, David Triana Robles, Mouhcine El Oualidi Charchmi, Ines Salamanca Estévez, and Noemí DeCastro-García, "Generative artificial intelligence and machine translators in spanish translation of early vulnerability cybersecurity alerts," *Applied Sciences*, vol. 15, no. 8, 2025.

[6] OpenAI, "ChatGPT 3.5," https://openai.com/en/, Último acceso: 17 de febrero de 2025.

[7] Microsoft, "Copilot," https://www.microsoft.com/en-us/microsoft-copilot, Último acceso: 17 de febrero de 2025.

[8] Gobierno de España, "ALIA - Artificial Inteligence," https://www.alia.gob.es/, Último acceso: 17 de febrero de 2025.

[9] INCIBE, "BBDD de vulnerabilidades," Último acceso: 6 de febrero de 2025. [Online]. Available: https://drive.google.com/file/d/1tJOP8R98DcEMCNh6whXkqTsXGFpoTYU2/view?usp=sharing

[10] Microsoft, "Bing translator," https://www.bing.com/translator?to=en&setlang=en, Último acceso: 17 de febrero de 2025.

[11] D. GmbH, "Deepl," https://www.deepl.com/en/translator, Último acceso: 17 de febrero de 2025.

[12] Google, "Google translate," https://translate.google.es/?hl=es&sl=fr&tl=es&op=translate, Último acceso: 5 de febrero de 2025.

[13] Microsoft, "Copilot for security," https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot, Último acceso: 17 de febrero de 2025.

[14] Aitor Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, José Javier Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, and M. Villegas, "Salamandra technical report," 2025, Último acceso: 17 de febrero de 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.08489

[15] MITRE Corporation, "Common Vulnerabilities and Exposures (CVE) - Introduction," 2021, Último acceso: 5 de febrero de 2025. [Online]. Available: https://cve.mitre.org/docs/cve-intro-handout.pdf

[16] P. Gupta, S. Shekhawat, and K. Kumar, "Unsupervised quality estimation without reference corpus for subtitle machine translation using word embeddings," in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 32–38, Último acceso: 18 de diciembre de 2024.

[17] INCIBE, "NVD Database Vulnerability Descriptions," 2019, Último acceso: 26 de noviembre de 2024. [Online]. Available: https://elrc-share.eu/repository/browse/descripciones-de-vulnerabilidades-de-la-bbdd-nvd/9dd01c02cfcc11e9913100155d026706700705608 07f458f8e50bd5273778d7a/

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 07 2016.

[19] F. Research, "Muse," https://github.com/facebookresearch/MUSE, Último acceso: 30 de septiembre de 2024.

Tabla IV
RESULTADOS DE LA PRUEBA DE *word embeddings* USANDO EL MODELO EN ESPAÑOL PREENTRENADO.

Traductor	Generador	Media	Desv. típica	Mediana	Máximo	Mínimo	Tiempo Medio
INCIBE	ChatGPT	0.6512	0.1789	0.6495	0.8727	0.2476	0.1923
	Copilot	0.6884	0.1893	0.7061	0.9176	0.2843	0.1888
	Salamandra	0.6356	0.1642	0.7000	0.8679	0.2722	0.2768
ChatGPT	ChatGPT	0.6669	0.1816	0.6847	0.9199	0.2623	0.2168
	Copilot	0.6935	0.1805	0.7111	0.9161	0.2618	0.1709
	Salamandra	0.6741	0.1516	0.7025	0.8873	0.3153	0.2181
Bing	ChatGPT	0.6449	0.1706	0.6531	0.8761	0.2648	0.1937
	Copilot	0.6533	0.1720	0.6438	0.8735	0.2720	0.1584
	Salamandra	0.6549	0.1720	0.7248	0.8577	0.2486	0.2248
DeepL	ChatGPT	0.6492	0.1764	0.6606	0.9062	0.2318	0.1783
	Copilot	0.6882	0.1904	0.7312	0.8907	0.2204	0.1621
	Salamandra	0.6563	0.1567	0.7064	0.9138	0.2632	0.1989
Google Translate	ChatGPT	0.6373	0.1771	0.6621	0.9130	0.2231	0.1786
	Copilot	0.6588	0.1875	0.6558	0.9201	0.2213	0.1593
	Salamandra	0.6378	0.1694	0.6925	0.8684	0.2798	0.1995
CFS	ChatGPT	0.6543	0.1706	0.6662	0.8723	0.2595	0.1706
	Copilot	0.6680	0.1742	0.6823	0.8918	0.2641	0.1526
	Salamandra	0.6638	0.1843	0.6751	0.8838	0.2307	0.2048
Copilot	ChatGPT	0.6395	0.1767	0.6613	0.8723	0.2450	0.1785
	Copilot	0.6919	0.1877	0.7179	0.9154	0.2576	0.1691
	Salamandra	0.6886	0.1474	0.7377	0.8700	0.2799	0.1920
CFS1	ChatGPT	0.6421	0.1731	0.6456	0.8961	0.2642	0.1728
	Copilot	0.7018	0.1893	0.7169	0.9391	0.2618	0.1780
	Salamandra	0.6351	0.1451	0.6723	0.8402	0.2499	0.1985
CFS2	ChatGPT	0.6738	0.1736	0.6844	0.8996	0.2677	0.1809
	Copilot	0.6807	0.1842	0.7204	0.9135	0.2618	0.1585
	Salamandra	0.6638	0.1642	0.7166	0.8784	0.2439	0.2214
Salamandra	ChatGPT	0.6787	0.1513	0.7188	0.8725	0.2989	0.2111
	Copilot	0.7084	0.1355	0.7261	0.8881	0.3417	0.2015
	Salamandra	0.6233	0.1526	0.6748	0.8434	0.2615	0.2168
Descripción original en inglés	ChatGPT	0.7191	0.1428	0.7548	0.8919	0.3114	0.2259
	Copilot	0.7705	0.1207	0.7871	0.9123	0.3715	0.1981
	Salamandra	0.6676	0.1312	0.6841	0.8918	0.3046	0.3018

Tabla V
COMPARACIÓN DE LAS DIFERENCIAS PROMEDIO ENTRE LA VALORACIÓN OBTENIDA POR LOS TÍTULOS EN FUNCIÓN DE LA DESCRIPCIÓN UTILIZADA.

Modelo	Generador	Descripción en inglés	Descripciones en español (Promedio)	Diferencia
1	ChatGPT	0.7899	0.72684	0.06306
	Copilot	0.8107	0.74476	0.06594
	Salamandra	0.7477	0.73869	0.00901
2	ChatGPT	0.8029	0.73566	0.06724
	Copilot	0.8288	0.75305	0.07575
	Salamandra	0.7637	0.7556	0.0081
3	ChatGPT	0.7191	0.65379	0.06531
	Copilot	0.7705	0.6833	0.0872
	Salamandra	0.6676	0.65333	0.01427

[20] P. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.
[21] S. Developers, "Scipy: Open source scientific tools for Python," Último acceso: 26 de noviembre de 2024. [Online]. Available: https://www.scipy.org/
[22] N. Developers, "Numpy: The fundamental package for scientific computing with Python," Último acceso: 3 de marzo de 2025. [Online]. Available: https://numpy.org/
[23] P. Developers, "Pandas: Python Data Analysis Library," Último acceso: 3 de marzo de 2025. [Online]. Available: https://pandas.pydata.org/

Sesión Investigación III: Ciberseguridad en IoT y entornos inteligentes

Estrategia de Agregación Ponderada para Sistemas IDS Federados en IoT

Luis Miguel García-Sáez* [1], Sergio Ruiz-Villafranca† [1],
José Roldán-Gómez† [2], Javier Carrillo-Mondéjar† [2],
José Luis Martínez† [1]

[1]Universidad de Castilla-La Mancha, C/ Investigación 2, Albacete 02071, España.
E-mail: {luism.garcia, sergio.rvillafranca, joseluis.martinez}@uclm.es

[2]Universidad de Zaragoza, C. María de Luna, 3, 50018 Zaragoza, España.
E-mail: {jroldan, jcarrillo}@unizar.es

* Corresponding author † Contributing author

***Resumen*—La rápida expansión y crecimiento de los entornos *Internet of Things* (*IoT*) e *Industrial Internet of Things* (*IIoT*) ha provocado un aumento en el número de ataques y riesgos en estos entornos, planteando nuevos desafíos de ciberseguridad que exigen sistemas de detección de intrusiones (*IDS*) más avanzados. Sin embargo, los *IDS* basados en *Machine Learning* (*ML*) centralizado afrontan problemas de escalabilidad, latencia y privacidad. En este ámbito, el aprendizaje federado (*FL*) ofrece un enfoque descentralizado que permite a múltiples nodos entrenar modelos de forma colaborativa sin exponer datos sensibles. Este trabajo presenta una propuesta de *IDS* federado basado en aprendizaje supervisado e introduce una estrategia de agregación que pondera dinámicamente las actualizaciones según la calidad y la incertidumbre de los datos locales. La arquitectura planteada es evaluada a través de diferentes *datasets* de tráfico *IoT/IIoT* orientados a ciberseguridad y ampliamente utilizados en estos entornos, mostrando un rendimiento equiparable e incluso superior a los métodos centralizados, con un *F1-Score* promedio que oscila entre 0.98 - 0.99 para las pruebas realizadas. Estos resultados evidencian la capacidad y potencial de *FL* en la detección de intrusiones, aprovechando de manera eficaz las ventajas de escalabilidad y privacidad que ofrece.**

***Index Terms*—Ciberseguridad, Internet of Things, Federated Learning, Intrusion Detection System, Machine Learning**

Tipo de contribución: *Investigación original*

I. Introducción

La constante evolución de las redes y la creciente adopción de entornos *Internet of Things* (*IoT*) e *Industrial Internet of Things* (*IIoT*) [1] han propiciado que millones de dispositivos conectados generen y transmitan grandes cantidades de datos, transformando la forma de interacción con el entorno en ámbitos tan diversos como la salud, el transporte o los hogares inteligentes [2]. Se estima que para el año 2030 habrá más de 30 mil millones de dispositivos *IoT* interconectados alrededor del mundo [3]. Esto conlleva un aumento de más del 50 % con respecto al número de dispositivos *IoT* existentes actualmente, así como un aumento en la complejidad de las infraestructuras. Esta multiplicidad de dispositivos y la diversidad de fabricantes y protocolos incrementan la complejidad de la red y, por ende, amplían sustancialmente la superficie de ataque.

En el ámbito industrial, el *IIoT* lleva estos conceptos un paso más allá al aplicarlos en sectores críticos y de gran escala dentro del marco de la Industria 4.0 y en transición hacia la Industria 5.0 [4], integrando inteligencia artificial, automatización y robótica para optimizar procesos en tiempo real. Sin embargo, la propia naturaleza de *IIoT* introduce un grado de criticidad que supera al de *IoT*, pues un simple fallo puede llegar a suponer grandes pérdidas económicas [5].

En este contexto surge el concepto de *Multi-access Edge Computing* (*MEC*), un paradigma que incorpora numerosas ventajas capaces de ser aprovechadas en estos entornos. *MEC* plantea la posibilidad de trasladar capacidades de cómputo y almacenamiento desde la nube hacia nodos más cercanos de los dispositivos finales [6], lo que permite aprovechar mejoras en latencia y ancho de banda, a la vez que reducir la superficie potencial de ataque asociada a la comunicación con la nube. Sin embargo, el crecimiento exponencial en los entornos *IoT*/*IIoT* ha venido acompañado de la aparición de nuevas amenazas [7], sumado al aumento considerable del número de ataques dirigidos a infraestructuras *IoT*.

Frente a estos desafíos de seguridad y a fin de mitigar los riesgos asociados al aumento de ataques, se han propuesto diversas soluciones centradas en la detección y mitigación temprana de actividades maliciosas en la red. Dentro de este ámbito, los Sistemas de Detección de Intrusiones (*IDS*) han surgido como una pieza importante para identificar comportamientos anómalos, bloqueando potenciales intrusos o incidentes [8]. En sus inicios, estos sistemas se basaban principalmente en reglas y firmas predefinidas, eficaces para detectar amenazas conocidas pero menos ágiles ante ataques novedosos o variantes de patrones maliciosos preexistentes. Con el tiempo, se han explorado nuevas propuestas, como los *IDS* basados en *Machine Learning* (*ML*), que suponen un enfoque más sofisticado basado en la inteligencia artificial y el aprendizaje automático [9].

En este contexto, los *IDS* basados en aprendizaje federado (*FL*) [10] emergen como una vertiente de investigación prometedora que preserva la privacidad de los datos de entrenamiento mientras aprovecha la colaboración en entornos distribuidos. A diferencia de los métodos tradicionales de *ML* que centralizan datos, este enfoque permite que múltiples nodos entrenen un modelo conjunto intercambiando solo los parámetros aprendidos con un servidor central [11], existiendo incluso propuestas descentralizadas que eliminan dicho servi-

dor y aumentan la robustez del sistema [12]. Todo ello resulta particularmente adecuado para entornos *IoT* con restricciones de privacidad o alta heterogeneidad.

La combinación de *FL* junto a la tecnología *MEC* en entornos *IoT*, refleja una solución que ofrece un equilibrio entre rendimiento, escalabilidad y privacidad, aprovechando la presencia de recursos computacionales cada vez más potentes en el paradigma *MEC*. Esto convierte a los *IDS* basados en *FL* (*FL-IDS*) en propuestas altamente prometedoras para los escenarios planteados, al ser capaces de abordar la diversidad de amenazas y la dinámica de las redes distribuidas sin sacrificar el rendimiento de las aplicaciones y servicios críticos.

Sin embargo, la mayoría de trabajos e investigaciones recientes se centran en enfoques basados en aprendizaje no supervisado, dado que la agregación de gradientes se integra de forma muy natural en el paradigma federado [13]. Sin embargo, aunque existen algunas propuestas que emplean algoritmos concretos de aprendizaje supervisado [14], este predominio deja abierta una ventana de investigación: explorar cómo se beneficiarían los esquemas federados del rendimiento, la interpretabilidad y la eficiencia probada de algoritmos supervisados clásicos basados en árboles, en contextos heterogéneos propios de entornos *IoT/IIoT*.

Por ello, en este trabajo proponemos una arquitectura modular y adaptativa de *FL-IDS* basada en aprendizaje supervisado, capaz de integrar múltiples algoritmos de clasificación basados en árboles para adaptarse a los entornos *IoT* tradicionales e industriales, así como a la diversidad de ataques que abundan en estos entornos. Asimismo, bajo esta arquitectura introducimos una estrategia de agregación dinámica, *FedWLA*, que considera la calidad y la incertidumbre de los datos locales en cada nodo, permitiendo afinar y adaptar el modelo global en base a los desempeños locales. De acuerdo a lo expuesto, este trabajo presenta las siguientes contribuciones:

- Se plantea un *FL-IDS* donde los modelos se entrenan de forma distribuida en nodos *MEC* cercanos a los dispositivos *IoT* finales que generan los datos, aprovechando las ventajas de la tecnología *MEC* y mejorando el grado de preservación de la privacidad de los datos.
- Se adaptan y evalúan al contexto federado, algoritmos centralizados de aprendizaje automático basados en árboles, como *Catboost*, *XGBoost*, *LightGBM*, entre otros [15], aprovechando sus ventajas de eficacia y eficiencia en escenarios heterogéneos de *IoT/IIoT*.
- Propuesta de una estrategia de agregación federada (*FedWLA*). Se introduce un mecanismo de ponderación dinámico que considera métricas locales (calidad de los datos, nivel de incertidumbre de las predicciones y volumen de muestras), permitiendo establecer una ponderación variable independiente para cada nodo de la red.
- Se realiza un estudio comparativo del rendimiento de los modelos federados y su versión no federada a través de los *datasets X-IIoTID*, *WUSTL-IIOT-2021* e *IoT-23*. Se analiza en detalle la eficacia de cada algoritmo en la arquitectura propuesta, comparándola con su ejecución centralizada tradicional y con diferentes estrategias de agregación disponibles en la literatura, tales como *FedAvg*, *FedProx*, etc. incluyendo la estrategia propuesta *FedWLA*.

El resto del trabajo se organiza de la siguiente manera. En la Sección II se revisan los trabajos relacionados, desde propuestas tradicionales centralizadas hasta propuestas más recientes basadas en *FL*. En la Sección III se detalla la arquitectura propuesta, describiendo sus componentes, algoritmos seleccionados y la estrategia de agregación federada propuesta. A continuación, en la Sección IV se presenta el entorno experimental y la metodología de evaluación, incluyendo los *datasets* utilizados. La Sección V analiza los resultados obtenidos y su relevancia, comparándolos con enfoques tradicionales y valorando la eficacia de cada estrategia de federación. Finalmente, en la Sección VI se exponen las conclusiones generales y se proponen posibles líneas de trabajo futuro.

II. Trabajos Relacionados

En esta sección se repasa el estado del arte de los sistemas de detección de intrusiones, desde aquellos basados en modelos clásicos de aprendizaje automático, hasta las iniciativas más recientes que aplican de manera innovadora *FL*.

Diversos trabajos han abordado la detección de intrusiones en redes *IoT* utilizando técnicas clásicas de *ML*. Por ejemplo, Awotunde et al. [16] proponen un modelo basado en clasificadores de árboles ensamblados (*ensemble-tree*), como *XGBoost*, *Random Forest* (*RF*) y *AdaBoost*, aplicado sobre el conjunto de datos *TON_IoT*, destacando especialmente el rendimiento de *XGBoost* con un *accuracy* cercano al 99.9 %. En la misma línea, Mohy-eddine et al. [17] utilizan un enfoque ensamblado basado en *RF*, aplicando *Isolation Forest* para detección de anomalías y *coeficiente de correlación de Pearson* para la selección de características, alcanzando un *accuracy* del 99.99 % sobre *Bot-IoT* y del 99.12 % en *wustl_iiot_2021*. Por otro lado, Altulaihan et al. [18] centran su propuesta en la detección de ataques *DoS* en redes *IoT* mediante un *IDS* basado en detección de anomalías, comparando el desempeño de cuatro clasificadores supervisados (*Decision Tree* (*DT*), *RF*, *kNN* y *SVM*) entrenados con el conjunto de datos *IoTID20*, siendo *DT* y *RF* los mejores, con un *accuracy* cercano al 100 %.

Sin embargo, a pesar de los excelentes resultados obtenidos por estos métodos tradicionales, propuestas recientes han explorado la integración de *FL* en los *IDS* con el objetivo de alcanzar soluciones distribuidas más escalables que logren preservar la privacidad de los datos empleados en el entrenamiento de los modelos de *ML*.

Belenguer et al. presentan *GöwFed* [19], un *FL-IDS* basado en redes neuronales cuyo objetivo es identificar amenazas en redes *IoT* preservando la privacidad. La arquitectura incorpora 6 capas ocultas totalmente conectadas con tamaños descendentes (128, 64, 64, 32, 32 y 2 neuronas) y una capa de *dropout* para regularización, utilizando *FedAvg* como técnica de agregación. En evaluaciones con el *dataset TON_IoT* alcanza un *accuracy* del 93 % frente al 98.9 % del modelo centralizado equivalente, mientras que una segunda versión con un mecanismo de atención basado en selección de nodos según su desempeño, logra resultados similares seleccionando el 80 % de los mejores agentes, con potencial particular para escenarios heterogéneos (*non-IID*) donde ayudaría a evitar contribuciones negativas de nodos maliciosos.

En la propuesta [20], los autores plantean *Fed-ANIDS*, un sistema distribuido de detección de intrusiones en redes basado en *FL* y detección de anomalías mediante *autoencoders simples* (*AE*), *autoencoders variacionales* (*VAE*) y *autoencoders adversariales (AAE)*, modelos que han demostrado gran efectividad para la identificación de patrones complejos en el tráfico de red [21]. El sistema emplea específicamente la estrategia federada *FedProx*, que generaliza y añade un término de regularización sobre el conocido método *FedAvg*, con el fin de mejorar la estabilidad y el rendimiento en escenarios heterogéneos. En sus experimentos, realizados sobre los *datasets USTC-TFC2016, CIC-IDS2017* y *CSE-CIC-IDS2018, Fed-ANIDS* logra un rendimiento superior a otros modelos basados en Redes Generativas Antagónicas (*GAN*) y muy similar, e incluso hasta un 20 % superior en algunos casos, al análisis centralizado realizado sobre los *datasets*. Concretamente, utilizando *FedProx* y *AAE* sobre el *dataset USTC-TFC2016*, obtienen un *accuracy* máximo del 99.95 % y un *F1-score* del 99.94 %. En *CIC-IDS2017*, el mejor rendimiento se consigue con *AE*, alcanzando un *accuracy* del 93.54 %. Finalmente, sobre *CSE-CIC-IDS2018*, empleando *VAE* obtienen un *accuracy* del 94.48 %, con un *F1-score* del 90.64 %.

Los autores de [22] proponen un *FL-IDS* denominado *Dynamic weighted Aggregation Federated Learning* (*DAFL*), orientado especialmente a entornos con datos *non-IID* (*non-Independent and Identically Distributed*). Como modelo base de aprendizaje automático, utilizan una Red Neuronal Convolucional (*CNN*) adaptada específicamente para reconocer patrones en tráfico de red transformado en imágenes bidimensionales. La principal innovación en su propuesta reside en la estrategia de agregación federada. Utilizan una variante dinámica de *FedAvg* que filtra continuamente los modelos locales con bajo rendimiento mediante un umbral de precisión predefinido, además de calcular pesos dinámicos basados en el desempeño individual de cada modelo local, priorizando así aquellos con mejores resultados. Experimentalmente, utilizando el *dataset CSE-CIC-IDS2018*, *DAFL* consigue un *accuracy* superior al 94 % con configuraciones variadas de clientes (3, 5 y 7 nodos), frente al 92,51 % obtenido por la versión federada estándar (*FedAvg*) y muy próximo al 94.44 % del modelo centralizado equivalente.

Rashid et al. [23] proponen un enfoque basado en *FL* para detección de intrusiones en redes *IIoT*, evaluando dos tipos de modelos de redes neuronales: *CNN* y Redes Neuronales Recurrentes tipo *LSTM* (*RNN-LSTM*). La estrategia de agregación federada empleada es *FedAvg*. Para la evaluación, se emplea el conjunto de datos *Edge-IIoTset*, diseñado específicamente para escenarios realistas de *IIoT*, evaluando dos distribuciones de datos (*IID* y *non-IID*). Tras 50 rondas de entrenamiento federado, el modelo basado en *CNN* obtuvo un *accuracy* global del 91.13 % en datos *IID* y 90.73 % en datos *non-IID*, mientras que el modelo *RNN* alcanzó mejores resultados, logrando una precisión global máxima del 92.37 % en *IID* y 91.87 % en distribución *non-IID*. Estos resultados se aproximan considerablemente a los obtenidos mediante entrenamiento centralizado, donde el *accuracy* fue del 93.92 %.

Friha et al. [24] proponen *2DF-IDS*, un *FL-IDS* descentralizado que utiliza *MLP*. El modelo neuronal está formado por tres capas ocultas (128, 64 y 32 neuronas respectivamente) y emplea una versión descentralizada de *FedAvg* para la agregación de parámetros, evitando la dependencia del servidor central. Experimentalmente, sobre el *dataset Edge-IIoTset*, *2DF-IDS* alcanza un *accuracy* de 94.37 %, resultado equivalente al obtenido por el modelo centralizado y superior a la versión federada centralizada empleando *FedAvg*, presentando una alternativa descentralizada más robusta (al eliminar el servidor central) y con mayor rendimiento que el planteamiento centralizado.

Tras revisar las propuestas existentes en la literatura, se aprecia una marcada diferencia entre los enfoques elegidos según sean *FL-IDS* o *ML-IDS*. Tradicionalmente, las propuestas centralizadas optan por algoritmos clásicos supervisados como *RF*, *DT* o modelos basados en árboles ensamblados (p.ej., *XGBoost*). Por otro lado, la mayoría de los *FL-IDS* recientes recurren a redes neuronales o métodos no supervisados como los *autoencoders*, posiblemente debido a la facilidad que ofrecen estos modelos para gestionar heterogeneidad y datos parcialmente etiquetados en escenarios distribuidos. No obstante, este predominio abre un horizonte de investigación centrado en analizar cómo el esquema federado puede aprovechar las características y ventajas de estos algoritmos en entornos *non-IID*, característicos de aplicaciones *IoT*.

Nuestro trabajo aborda esta vía de investigación combinando la escalabilidad y privacidad inherentes de *FL* con la eficiencia y capacidad de clasificación en datos tabulares (superior a la de las redes neuronales) de los algoritmos supervisados basados en árboles. A diferencia de las propuestas analizadas, enfocadas principalmente en redes neuronales o enfoques no supervisados, la arquitectura propuesta es capaz de integrar y ajustarse de forma modular a modelos como *DT*, *RF*, *CatBoost*, *LightGBM* o *XGBoost* dentro de un esquema federado que, además, introduce una nueva estrategia de agregación (*FedWLA*) que permite un ajuste local más preciso en cada nodo, mejorando la adaptación a escenarios heterogéneos. Asimismo, la propuesta no se limita únicamente a la detección binaria de intrusiones, sino que también permite identificar ataques específicos mediante clasificación multinivel, ofreciendo una mayor granularidad respecto a algunos trabajos previos.

III. Diseño del Sistema

Esta sección describe la arquitectura del *FL-IDS* planteado, incluyendo las capas que componen dicha arquitectura. Se detalla cómo se integran los clasificadores basados en árboles y el desarrollo de una estrategia de agregación propia denominada *FedWLA*, justificando su elección y funcionamiento.

III-A. Algoritmos Federados

El aprendizaje federado tradicionalmente implica compartir actualizaciones de parámetros o gradientes entre clientes para obtener un modelo global común. Sin embargo, esta metodología es natural y eficaz en modelos paramétricos continuos, como las redes neuronales, donde el proceso de agregación por promedio tiene un sentido matemático directo. En contraste, los algoritmos basados en árboles presentan ciertas particularidades que hacen que la agregación directa de parámetros mediante promedios de pesos o gradientes sea

menos efectiva, ya que estos modelos están formados por árboles generados a partir de decisiones binarias o reglas condicionales en lugar de pesos numéricos continuos que se ajustan de manera incremental.

La estructura interna de los árboles de decisión y los métodos ensamblados basados en árboles no permite una combinación simple de parámetros o pesos entre modelos locales distintos a través de las estrategias típicas de *FL*, como *FedAvg*, *FedProx* o *FedOpt*. Esto es debido a que estos modelos no poseen parámetros que puedan ser promediados linealmente de forma significativa. Para superar esta limitación inherente, en este trabajo optamos por una aproximación alternativa basada en la agregación federada de hiperparámetros del modelo. Esta aproximación permite aprovechar plenamente la capacidad predictiva individual de cada modelo local, al mismo tiempo que posibilita generar una solución global coherente mediante la combinación de configuraciones óptimas obtenidas en nodos individuales. Tras cada ronda de entrenamiento, los clientes comparten con el servidor central la configuración de hiperparámetros utilizada junto a su rendimiento asociado, y el servidor central aplica la estrategia de agregación seleccionada para cada caso, como *FedAvg*, *FedOpt*, entre otras. Este enfoque permite explorar simultáneamente diferentes regiones del espacio de búsqueda de hiperparámetros, acelerando así la convergencia hacia una configuración óptima global. De esta manera y tras revisar los algoritmos óptimos más utilizados en la literatura [25], [26], se ha optado por seleccionar los siguientes: *DT*, *RF*, *Catboost*, *LightGBM* y *XGBoost* [27].

III-B. Adaptación Local Ponderada Federada (FedWLA)

En *FL*, la estrategia utilizada para agregar los parámetros provenientes de cada cliente es un aspecto fundamental que determina la efectividad del modelo global. En este contexto, diversas estrategias han sido propuestas y ampliamente estudiadas, siendo las más destacadas las ya mencionadas *FedAvg*, *FedProx* y *FedOpt*:

- **FedAvg**: Es una de las estrategias más conocidas debido a su simplicidad y eficiencia. Su principio básico consiste en realizar una media ponderada de los parámetros locales enviados por los clientes, asignando pesos proporcionales al número de muestras de cada uno (que habitualmente suele ser el mismo). Es una estrategia que destaca por su facilidad de implementación y bajo costo computacional [28].
- **FedProx**: Constituye una evolución respecto a *FedAvg*, incorporando un término adicional de regularización proximal que limita las actualizaciones locales excesivamente divergentes. Este término adicional ayuda a estabilizar el entrenamiento global, especialmente cuando existen diferencias significativas entre los *datasets* de los distintos clientes [29].
- **FedOpt**: Representa otra variante que utiliza optimizadores avanzados en la fase de agregación global, tales como *Adam*, *SGD* con *momentum*, o *Adagrad*, con el objetivo de acelerar y mejorar la convergencia global [30].

A pesar de que estas estrategias han sido ampliamente adoptadas en numerosos escenarios, su aplicación a entornos con datos *non-IID* puede llegar a comprometer su desempeño. Es por ello que esta consideración impulsa el desarrollo de una estrategia propia denominada *Adaptación Local Ponderada Federada* (*FedWLA*).

FedWLA propone un enfoque híbrido que combina una ponderación dinámica global basada en características específicas como la incertidumbre de los modelos locales, la calidad intrínseca del *dataset* en cada cliente y el tamaño del mismo. La incertidumbre en las predicciones locales proporciona información sobre la confianza del modelo en cada nodo. Puede interpretarse como la dificultad del modelo para discernir entre patrones normales y de ataque en cada cliente, por lo que una menor incertidumbre indica que el modelo local ha encontrado patrones más representativos, recibiendo mayor peso en la agregación global.

Para estimar la incertidumbre en la estrategia propuesta, se utiliza la entropía promedio de las predicciones realizadas por el modelo local de cada cliente. Concretamente, dada una muestra x del conjunto de datos de validación local X_{local} el modelo entrenado en el cliente genera una distribución de probabilidad sobre cada una de las C clases posibles:

$$p_i(c) = P(Y = c \mid X_i), \quad \text{con } c \in \{1, 2, \ldots, C\} \tag{1}$$

Donde:

- $p_i(c)$ representa la probabilidad predicha por el modelo para la clase c en la muestra i.
- C es el número total de clases del problema.
- X_i es la muestra específica sobre la cual se está haciendo la predicción.

La incertidumbre asociada a esta predicción (la entropía de *Shannon* de las probabilidades de clasificación) se define mediante la siguiente expresión:

$$H(p_i) = -\sum_{c=1}^{C} p_i(c) \cdot \log_2(p_i(c)) \tag{2}$$

Para obtener una estimación general de la incertidumbre en todo el *dataset* local compuesto por N muestras de test, se calcula la incertidumbre promedio de todas las predicciones de la siguiente manera:

$$\text{uncertainty}_i = \frac{1}{N} \sum_{i=1}^{N} H(p_i) \tag{3}$$

Por otro lado, el índice de calidad del *dataset* local, definido inversamente por la desviación estándar de la distribución de clases, permite dar prioridad a los clientes que cuentan con *datasets* más equilibrados. En escenarios reales de *IoT* e *IIoT*, estos *datasets* suelen estar fuertemente desbalanceados al ser entornos altamente heterogéneos y en constante cambio, lo que agrega gran importancia al equilibrio local del *dataset* para evitar sesgos en el modelo global.

$$\text{data_quality} = \frac{1}{\text{std_distribution} + 1 \times 10^{-6}} \tag{4}$$

Finalmente, incorporar la cantidad de ejemplos de cada cliente permite otorgar más peso a aquellos que cuentan con

mayor volumen de datos. De esta forma, la fórmula que define el peso (w_i) para cada *cliente*$_i$ en la estrategia es:

$$w_i = \text{data_quality}_i \times \frac{1}{(\text{uncertainty}_i + 10^{-6})} \times \text{num_examples}_i \quad (5)$$

Además, una vez que cada cliente recibe los parámetros globales del servidor, estos parámetros no son adoptados directamente. En su lugar, se realiza una búsqueda local limitada alrededor de estos hiperparámetros globales mediante un *random search* (con un número reducido de iteraciones). El objetivo es permitir la personalización en cada nodo para capturar patrones específicos locales y reducir la probabilidad de que el modelo federado converja hacia parámetros subóptimos globales, explorando localmente alternativas que puedan proporcionar mejores resultados para contextos específicos. Por otro lado, aunque en esta propuesta *FedWLA* se evalúa bajo ejemplos concretos, la estrategia es agnóstica al tipo de modelo y puede ser adaptada para funcionar con diferentes algoritmos de aprendizaje automático manteniendo su carácter.

III-C. Arquitectura de FL-IDS Propuesta

En base a lo descrito, proponemos una arquitectura de *FL-IDS* que se aprecia en la Figura 1, basada en la combinación de tres capas principales: *Cloud*, *MEC* e *IoT*, optimizando el uso de recursos computacionales distribuidos y garantizando la privacidad de los datos locales generados por los dispositivos.

Capa *Cloud*: En esta capa se sitúa un nodo centralizado que actúa como servidor federado, encargado de realizar los cálculos globales y gestionar la política de agregación *FedWLA*. Su función es recibir las actualizaciones locales provenientes de cada nodo de la capa MEC, realizar la agregación ponderada y distribuir posteriormente un modelo global actualizado entre el resto de nodos de las estaciones *MEC*.

Capa *MEC*: Cada nodo *MEC* gestiona una red local específica compuesta por múltiples dispositivos *IoT*. En estos nodos se implementan modelos locales independientes basados en algoritmos tradicionales de *ML* mencionados. Estos modelos se entrenan utilizando datos generados únicamente dentro de su red local específica.

Capa *IoT*: La capa inferior incluye dispositivos como sensores y actuadores, conectados directamente a cada red local administrada por su nodo MEC correspondiente. Estos dispositivos generan continuamente datos operativos y de red que son preprocesados localmente, asegurando baja latencia y privacidad de los datos generados.

IV. Experimentación

En esta sección se define el entorno de pruebas utilizado para evaluar la propuesta, detallando los *datasets*, las métricas de rendimiento y los parámetros experimentales. Además, se especifica cómo se distribuyen los nodos y se prepara la infraestructura de *FL*.

IV-A. Datasets

Para el entorno de pruebas se han seleccionado tres conjuntos de datos representativos que abarcan diversas características y escenarios del tráfico en redes *IoT* e *IIoT*:

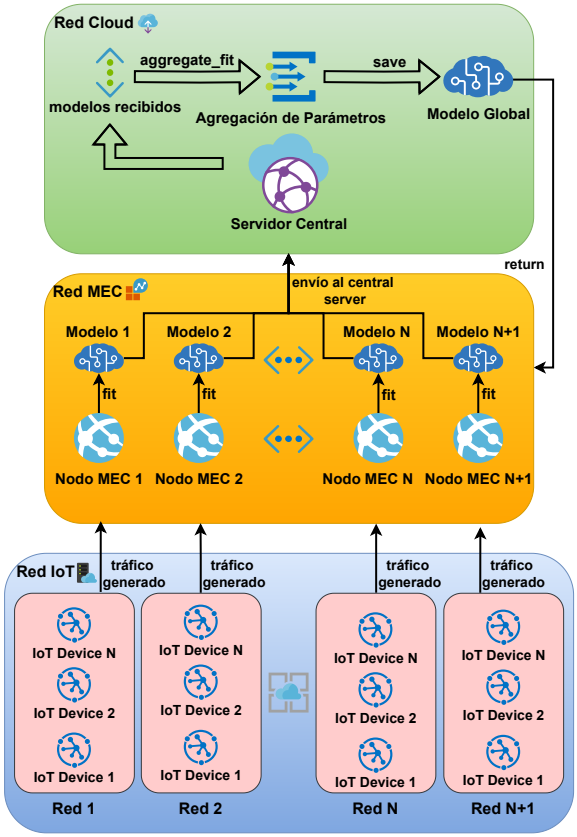


Figura 1. Arquitectura *FL-IDS* propuesta

- *X-IIoTID* es un *dataset* centrado en *IIoT*. Los datos se organizan en varias columnas que describen características estadísticas del tráfico, información de red y campos de protocolo. Su independencia de la tecnología de conexión y de los dispositivos lo hace altamente versátil, lo que permite la evaluación multiclasificación de sistemas *IDS* [31].
- *IoT-23* agrupa 23 escenarios con tráfico benigno y malicioso procedente de dispositivos *IoT*. Proporciona capturas *pcap* y metadatos, cubriendo ataques como *Mirai* o *Gafgyt* y facilitando la labor de análisis en estos entornos [32].
- *WUSTL-IIOT-2021* se basa en tráfico real capturado en entornos industriales con múltiples protocolos (*Modbus*, *OPC UA*, *MQTT*, entre otros). Está etiquetado a nivel de flujo de red, lo que facilita identificar los tipos de ataque e incluye tanto sesiones benignas como secuencias de simulación de ataques [33].

IV-B. Métricas

Para evaluar adecuadamente la eficacia del modelo en un contexto de clasificación múltiple de ataques, es fundamental emplear diversas métricas que proporcionan una visión integral del rendimiento. Estas métricas se derivan principalmente de la matriz de confusión, construida con los resultados generados por el modelo al clasificar cada instancia del tráfico

de red. La matriz ofrece información clave mediante cuatro indicadores principales: verdaderos positivos (*TP*), verdaderos negativos *(TN)*, falsos positivos (*FP*) y falsos negativos (*FN*), fundamentales para analizar el rendimiento del sistema en entornos distribuidos. Definimos el significado de dichos términos, en el contexto de la propuesta planteada, de la siguiente manera:

- Verdaderos Positivos (*TP*): Instancias de la clase positiva correctamente predichas.
- Verdaderos Negativos (*TN*): Instancias de la clase negativa correctamente predichas, es decir, aquellas que el modelo clasifica correctamente como no pertenecientes a una categoría específica de ataque
- Falsos Positivos (*FP*): Instancias negativas clasificadas erróneamente como positivas.
- Falsos Negativos (*FN*): Instancias positivas clasificadas erróneamente como negativas.

En base a estos indicadores clave, existen diversas métricas, ampliamente utilizadas en el ámbito del aprendizaje automático, para evaluar el rendimiento de los modelos de clasificación, tales como *accuracy*, *precision*, *recall*, o *F1-Score*. De entre ellas, empleamos el *F1-Score* como principal métrica de rendimiento, al ser aquella con mayor capacidad para captar simultáneamente la capacidad del modelo de identificar ataques sin incrementar el número de *FP* de forma desproporcionada, mostrando una visión global más equilibrada que evita penalizar demasiado los errores cuando la distribución de clases es muy asimétrica.

F1-Score. Es la media armónica entre *precision* y *recall*, proporcionando un equilibrio entre ambas métricas. Es realmente útil en *datasets* heterogéneos donde no hay un balance entre las clases.

$$\text{F1-Score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{6}$$

IV-C. Pruebas

El entorno de experimentación se compone de un total de 10 nodos participantes, lo que supone un número realista [34] para la distribución de procesos de *FL* en una red *IoT/IIoT*. Para cada algoritmo de *ML* considerado, se han utilizado diferentes estrategias de federación. En concreto, para los algoritmos *RF*, *DT*, *LightGBM* y *Catboost*, se ha evaluado la efectividad de las estrategias *FedAvg*, *FedProx* y *FedOpt*. Sin embargo, en el caso específico de *XGBoost*, debido a sus particularidades de entrenamiento, se ha llevado a cabo una adaptación específica propia de la estrategia *FedAvg* y se han incorporado adicionalmente dos estrategias especializadas: *FedXgbBagging* y *FedXgbCyclic*. Además de estas estrategias, para todos los algoritmos se ha evaluado la efectividad de la estrategia propuesta *FedWLA*. Para asegurar la validez y representatividad estadística de los resultados obtenidos, cada escenario ha sido ejecutado durante 15 iteraciones independientes, obteniendo como resultado métricas promedio para asegurar su representatividad. Finalmente, cada iteración experimental ha sido ejecutada durante 10 rondas completas de entrenamiento.

V. Resultados

A continuación, en las Figuras 2, 3 y 4 se muestran los resultados obtenidos (por algoritmo y estrategia federada de agregación) para los *datasets* seleccionados, empleando el *F1-Score* como métrica principal para evaluar el rendimiento.

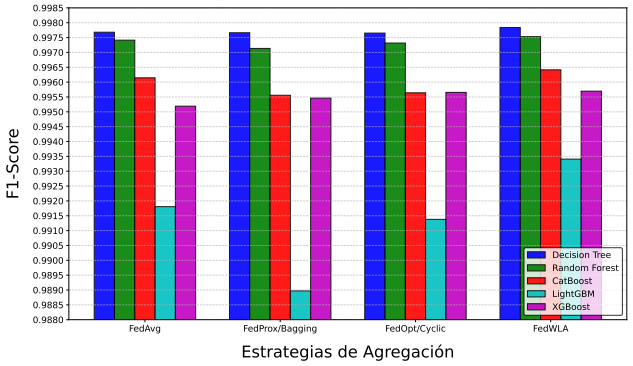


Figura 2. Comparación del *F1-Score* para el *dataset X-IIoTID*

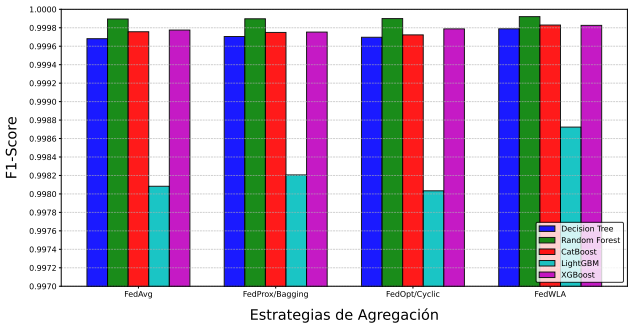


Figura 3. Comparación del *F1-Score* para el *dataset WUSTL-IIOT-2021*

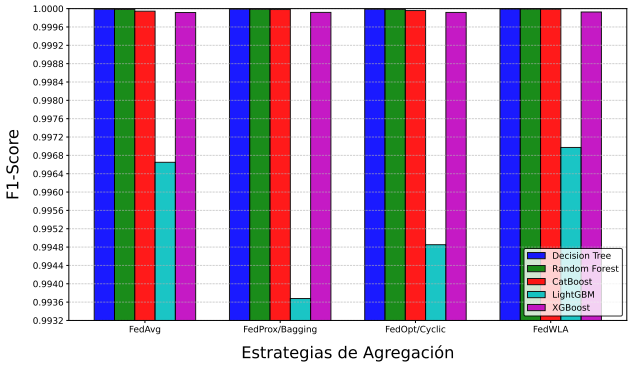


Figura 4. Comparación del *F1-Score* para el *dataset IoT-23*

Los resultados obtenidos muestran que *FedWLA*, la variante propuesta en este trabajo, logra el mejor *F1-Score* en la mayoría de los experimentos, superando a *FedAvg*, *FedProx* y *FedOpt*, y logrando, en el resto de ocasiones, un rendimiento equivalente al de estas. Esto muestra la importancia de ponderar de forma personalizada la contribución de cada nodo y permitir ajustes locales tras recibir la versión del modelo global, previamente agregada. Además, al no realizar una distribución uniforme de los datos del *dataset* (*non-IID*), la

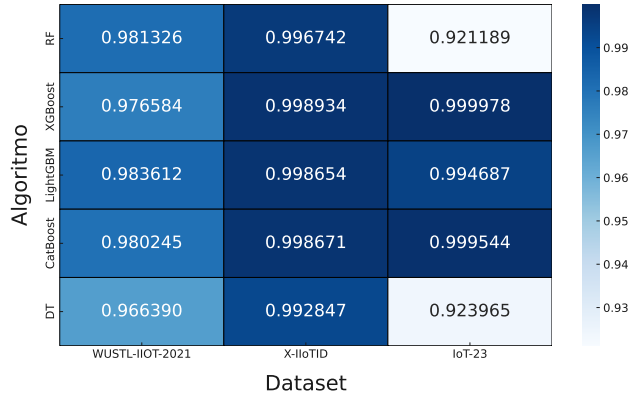


Figura 5. Resultados de *F1-Score* de las versiones sin federar

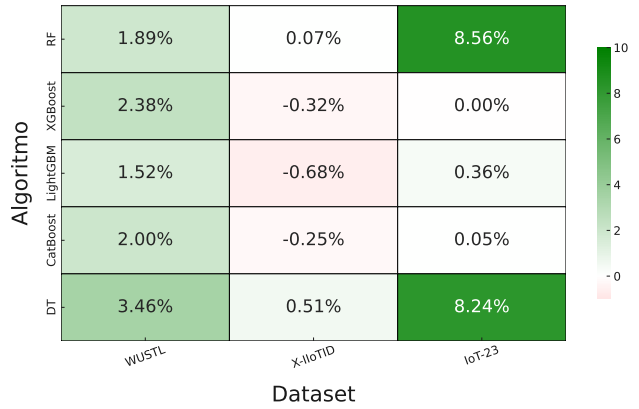


Figura 6. Mejora del *F1-Score* de la propuesta sobre la versión tradicional

ponderación de 2 índices personalizados para cada nodo local puede contribuir a la mejora de *FedWLA* sobre el resto de las estrategias.

Por otro lado, en la Figura 5 disponemos de los resultados obtenidos por las respectivas versiones sin federar de estos algoritmos. Si comparamos estos resultados con los federados, podemos apreciar en la Figura 6 que los resultados son altamente similares a los obtenidos por su versión tradicional, aunque en determinados casos se llegan a obtener mejoras de hasta un 2 %, un 3 % o incluso un 8 % sobre estos. Estas mejoras pueden explicarse debido a:

- La heterogeneidad del entorno y de los *datasets*, lo que permite a los nodos identificar patrones propios de su "subentorno" y contribuir a la capacidad global del modelo.
- La federación puede ayudar a limitar el sobreajuste que se produciría si un único modelo dependiese en exceso de un conjunto de muestras poco representativo.
- Al sumar rondas de entrenamiento, el modelo global evoluciona de forma iterativa. Esto, en muchos casos, supera al modelo centralizado que se entrena "de una vez" sobre todo el *dataset*.

VI. Conclusiones y Trabajo Futuro

La propuesta planteada en este trabajo demuestra que el aprendizaje federado es una opción competitiva para la detección de intrusiones en redes *IoT/IIoT*. La combinación de *FL* junto a algoritmos basados en árboles comprueba, mediante pruebas con distintos conjuntos de datos (*IoT-23*, *WUSTL-IIoT*, *X-IIoTID*), que su versión federada es capaz de igualar o incluso superar el rendimiento centralizado con hasta un 8 % de mejora en *F1-Score* en algunos casos, en especial en entornos no homogéneos. Con ello, logramos aprovechar las fortalezas y ventajas de *FL* respecto a la privacidad de los datos sin ver perjudicado el rendimiento del modelo utilizado. Además, el despliegue sobre una infraestructura *MEC* refuerza la escalabilidad, al distribuir la carga de cómputo y reducir el riesgo de fallos críticos. Con ello, se consigue un modelo más robusto, capaz de adaptarse localmente a ataques específicos y, a la vez, beneficiarse de las ventajas de un enfoque colaborativo.

En cuanto a las posibles vías de trabajo futuro, la propuesta abre numerosas posibilidades de desarrollo, sentando una base para que avances que se apoyen en la arquitectura planteada, aborden nuevos desafíos de seguridad y extiendan su aplicabilidad a distintos escenarios y entornos. No obstante, estas vías de trabajo futuro vienen impulsadas por la presencia de ciertas limitaciones inherentes, donde destaca la dependencia de un servidor centralizado, la falta de mecanismos avanzados para proteger la privacidad de los parámetros intercambiados y la ausencia de sistemas específicos para detectar y mitigar el impacto de nodos maliciosos en el entrenamiento federado. Estas limitaciones motivan una serie de líneas de trabajo futuro, entre las que destacamos las siguientes:

- La adopción de métodos que permitan cifrar los parámetros comunicados por los modelos, como la privacidad diferencial o el cifrado homomórfico, así como la inclusión de mecanismos de detección de nodos maliciosos (para contrarrestar ataques de envenenamiento) que permitan reforzar aún más la seguridad y la integridad de los parámetros compartidos.
- La arquitectura propuesta contempla un servidor global que coordina la agregación. Explorar versiones *peer-to-peer* o con topologías descentralizadas eliminaría este punto único de fallo y, posiblemente, incrementaría la resiliencia, a costa de una lógica de federación más compleja.
- Visto el rendimiento de *FedWLA*, se abre la puerta a investigar métodos híbridos o más sofisticados (por ejemplo, el uso de *meta-learning* o heurísticas adaptativas) para optimizar, de forma conjunta y robusta, los hiperparámetros de los algoritmos basados en árboles dentro del contexto federado.
- La adopción de redes definidas por software (*SDN*) y tecnologías de comunicación de nueva generación (*5G/6G*) abre la puerta a una gestión más flexible y con mayor ancho de banda, permitiendo extender la propuesta a casos de uso como vehículos conectados (*IoV*) o ciudades inteligentes, entre otros.

Agradecimientos

Este trabajo ha sido financiado por la Universidad de Castilla-La Mancha mediante el contrato predoctoral, 2024-UNIVERS-12844, financiado por el Fondo Social Europeo Plus (FSE+), por la JCCM mediante el proyecto SBPLY/21/180501/000195 y mediante el proyecto I+D

PID2021-123627OB-C52, financiado por el MCIN y el Fondo Europeo de Desarrollo Regional: “una forma de hacer Europa”. Este trabajo también ha sido parcialmente apoyado por TED2021-131115A-I00 y PID2022-142332OA-I00, financiados por MCIN/AEI/10.13039/501100011033, por los fondos del Plan de Recuperación, Transformación y Resiliencia, por la Unión Europea (Next Generation), por el Instituto Nacional de Ciberseguridad de España (INCIBE) bajo *Proyectos Estratégicos de Ciberseguridad – CIBERSEGURIDAD EINA UNIZAR* –, y por el Departamento de Universidad, Industria e Innovación del Gobierno de Aragón bajo el *Programa de Proyectos Estratégicos de Grupos de Investigación* (grupo de investigación DisCo, ref. T21-23R).

Referencias

[1] J. M. Kizza, “Internet of Things (IoT): Growth, Challenges, and Security,” in *Guide to Computer Network Security*, J. M. Kizza, Ed. Cham: Springer International Publishing, 2024, pp. 557–573.

[2] C. Paul, A. Ganesh, and C. Sunitha, “An overview of IoT based smart homes,” in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, Jan. 2018, pp. 43–46.

[3] Lionel Sujay Vailshery, “IoT connections worldwide 2022-2033,” Nov. 2024. [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

[4] R. Tallat, A. Hawbani, X. Wang, A. Al-Dubai, L. Zhao, Z. Liu, G. Min, A. Y. Zomaya, and S. Hamood Alsamhi, “Navigating Industry 5.0: A Survey of Key Enabling Technologies, Trends, Challenges, and Opportunities,” *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 1080–1126, 2024, conference Name: IEEE Communications Surveys & Tutorials.

[5] J. Beerman, D. Berent, Z. Falter, and S. Bhunia, “A Review of Colonial Pipeline Ransomware Attack,” in *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*, May 2023, pp. 8–15.

[6] P. Ranaweera, A. D. Jurcut, and M. Liyanage, “Survey on Multi-Access Edge Computing Security and Privacy,” *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1078–1124, 2021, conference Name: IEEE Communications Surveys & Tutorials.

[7] Olukunle Oladipupo Amoo, Femi Osasona, Akoh Atadoga, Benjamin Samson Ayinla, Oluwatoyin Ajoke Farayola, and Temitayo Oluwaseun Abrahams, “Cybersecurity threats in the age of IoT: A review of protective measures,” *International Journal of Science and Research Archive*, vol. 11, no. 1, pp. 1304–1310, Feb. 2024.

[8] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, “Survey of intrusion detection systems: techniques, datasets and challenges,” *Cybersecurity*, vol. 2, no. 1, p. 20, Jul. 2019.

[9] M. Almseidin, M. Alzubi, S. Kovacs, and M. Alkasassbeh, “Evaluation of machine learning algorithms for intrusion detection system,” in *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, Sep. 2017, pp. 000 277–000 282, iSSN: 1949-0488.

[10] S. Agrawal, S. Sarkar, O. Aouedi, G. Yenduri, K. Piamrat, M. Alazab, S. Bhattacharya, P. K. R. Maddikunta, and T. R. Gadekallu, “Federated Learning for intrusion detection system: Concepts, challenges and future directions,” *Computer Communications*, vol. 195, pp. 346–361, Nov. 2022.

[11] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, “Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications,” *IEEE Access*, vol. 8, 2020.

[12] E. T. Martínez Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, “Decentralized Federated Learning: Fundamentals, State of the Art, Frameworks, Trends, and Challenges,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023, conference Name: IEEE Communications Surveys & Tutorials.

[13] B. Ghimire and D. B. Rawat, “Recent Advances on Federated Learning for Cybersecurity and Cybersecurity for Federated Learning for Internet of Things,” *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8229–8249, Jun. 2022, conference Name: IEEE Internet of Things Journal.

[14] T. Markovic, M. Leon, D. Buffoni, and S. Punnekkat, “Random Forest Based on Federated Learning for Intrusion Detection,” in *Artificial Intelligence Applications and Innovations*, I. Maglogiannis, L. Iliadis, J. Macintyre, and P. Cortez, Eds. Cham: Springer International Publishing, 2022, pp. 132–144.

[15] J. P. Ntayagabiri, Y. Bentaleb, J. Ndikumagenge, and H. El Makhtoum, “A Comparative Analysis of Supervised Machine Learning Algorithms for IoT Attack Detection and Classification,” *Journal of Computing Theories and Applications*, vol. 2, no. 3, pp. 395–409, Feb. 2025, num Pages: 15 Number: 3.

[16] J. B. Awotunde, S. O. Folorunso, A. L. Imoize, J. O. Odunuga, C.-C. Lee, C.-T. Li, and D.-T. Do, “An Ensemble Tree-Based Model for Intrusion Detection in Industrial Internet of Things Networks,” *Applied Sciences*, vol. 13, no. 4, p. 2479, Jan. 2023, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[17] M. Mohy-eddine, A. Guezzaz, S. Benkirane, and M. Azrour, “An effective intrusion detection approach based on ensemble learning for IIoT edge computing,” *Journal of Computer Virology and Hacking Techniques*, vol. 19, no. 4, pp. 469–481, Nov. 2023.

[18] E. Altulaihan, M. A. Almaiah, and A. Aljughaiman, “Anomaly Detection IDS for Detecting DoS Attacks in IoT Networks Based on Machine Learning Algorithms,” *Sensors*, vol. 24, no. 2, p. 713, Jan. 2024, number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[19] A. Belenguer, J. A. Pascual, and J. Navaridas, “GöwFed: A novel federated network intrusion detection system,” *Journal of Network and Computer Applications*, vol. 217, p. 103653, Aug. 2023.

[20] M. J. Idrissi, H. Alami, A. El Mahdaouy, A. El Mekki, S. Oualil, Z. Yartaoui, and I. Berrada, “Fed-ANIDS: Federated learning for anomaly-based network intrusion detection systems,” *Expert Systems with Applications*, vol. 234, p. 121000, Dec. 2023.

[21] S. Schneider, D. Antensteiner, D. Soukup, and M. Scheutz, “Autoencoders - A Comparative Analysis in the Realm of Anomaly Detection,” 2022, pp. 1986–1992.

[22] J. Li, X. Tong, J. Liu, and L. Cheng, “An Efficient Federated Learning System for Network Intrusion Detection,” *IEEE Systems Journal*, vol. 17, no. 2, pp. 2455–2464, Jun. 2023, conference Name: IEEE Systems Journal.

[23] M. M. Rashid, S. U. Khan, F. Eusufzai, M. A. Redwan, S. R. Sabuj, and M. Elsharief, “A Federated Learning-Based Approach for Improving Intrusion Detection in Industrial Internet of Things Networks,” *Network*, vol. 3, no. 1, pp. 158–179, Mar. 2023, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[24] O. Friha, M. A. Ferrag, M. Benbouzid, T. Berghout, B. Kantarci, and K.-K. R. Choo, “2DF-IDS: Decentralized and differentially private federated learning-based intrusion detection system for industrial IoT,” *Computers & Security*, vol. 127, p. 103097, Apr. 2023.

[25] S. Choudhury and A. Bhowal, “Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection,” in *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, May 2015, pp. 89–95.

[26] S. Ruiz-Villafranca, J. Roldán-Gómez, J. M. C. Gómez, J. Carrillo-Mondéjar, and J. L. Martinez, “A TabPFN-based intrusion detection system for the industrial internet of things,” *The Journal of Supercomputing*, vol. 80, no. 14, pp. 20 080–20 117, Sep. 2024.

[27] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” *Information Fusion*, vol. 81, pp. 84–90, May 2022.

[28] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, “Reviewing Federated Learning Aggregation Algorithms; Strategies, Contributions, Limitations and Future Perspectives,” *Electronics*, vol. 12, no. 10, p. 2287, Jan. 2023, number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[29] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated Optimization in Heterogeneous Networks,” *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, Mar. 2020.

[30] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, “Adaptive Federated Optimization,” Sep. 2021, arXiv:2003.00295 [cs].

[31] M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, “X-IIoTID: A Connectivity-Agnostic and Device-Agnostic Intrusion Data Set for Industrial Internet of Things,” *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3962–3977, Mar. 2022, conference Name: IEEE Internet of Things Journal.

[32] S. Garcia, A. Parmisano, and M. J. Erquiaga, “IoT-23: A labeled dataset with malicious and benign IoT network traffic,” Jan. 2020.

[33] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, “WUSTL-IIOT-2021 Dataset for IIoT Cybersecurity Research,” Oct. 2021. [Online]. Available: https://www.cse.wustl.edu/~jain/iiot2/index.html

[34] S. Savazzi, M. Nicoli, and V. Rampa, “Federated Learning with Cooperating Devices: A Consensus Approach for Massive IoT Networks,” *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4641–4654, May 2020, arXiv:1912.13163 [eess].

Studying MQTT Malicious Traffic: Methods and Challenges

Mohammed Rashed
Universidad Carlos III de Madrid
Leganes, Spain
mrashed@inf.uc3m.es

Carlos García-Rubio
Universidad Carlos III de Madrid
Leganes, Spain
cgr@it.uc3m.es

Celeste Campo
Universidad Carlos III de Madrid
Leganes, Spain
celeste@it.uc3m.es

***Resumen*—Message Queuing Telemetry Transport (MQTT) is an application layer protocol widely used in Internet of Things (IoT) networks but which is also known for several security vulnerabilities. This paper proposes a machine learning (ML)-based approach to detect attacks on MQTT networks. Using network flow analysis and ML feature selection and ranking approaches, we provide a high-level description of our planned approach to detect brute-force attacks, Denial-of-Service (DoS), and malformed packet injection attacks on MQTT.**

***Index Terms*—MQTT, DoS, attack, network flow, malicious, argus**

Tipo de contribución: *Investigación en desarrollo (límite 8 páginas)*

I. Introduction

The IoT continues to grow exponentially, with billions of connected devices expected in operation in the next few years [1]. MQTT has emerged as one of the most widely used messaging protocols in IoT due to its lightweight publish/subscribe architecture and efficiency in low-bandwidth environments [2]. In MQTT, resource-constrained devices (publishers) are able to send sensor data to a central broker, which then distributes messages to subscribers based on topic subscriptions [3].

The decoupled design that MQTT supports is well-suited for diverse IoT applications ranging from smart homes to industrial automation [4]. However, the same features that make MQTT popular like simplicity and low overhead also introduce security challenges. MQTT networks often lack built-in security measures. For example, the protocol transmits data in plaintext by default (unless external TLS encryption is configured); even username/password authentication can be intercepted in this case. MQTT brokers and clients are vulnerable to various cyber threats. Common attack vectors include credential brute-force attacks on the broker, DoS attacks (such as flooding the broker with excessive messages or keeping connections open to exhaust resources), injection of malformed packets to exploit protocol parser weaknesses, and man-in-the-middle (MitM) interception of MQTT traffic [5], [6]. These attacks can lead to unauthorized access and data exfiltration or even disrupt IoT services. Traditional network intrusion detection systems (IDS) are not always effective in MQTT scenarios, since they may not recognize MQTT-specific patterns or handle the high volume of small IoT messages [7]. Ensuring security in MQTT deployments is therefore a critical concern. This requires solutions that take into account the specificity of the MQTT protocol and the resource-constraints in IoT networks.

To this end, our work aims to shed light on the security challenges that are MQTT-specific. An important aspect in facing these challenges is knowing the proper tools and data on which any detection system is based. Thus, in this work we focus on identifying the adequate tools and datasets that are oriented towards MQTT.

- We address the different MQTT attack scenarios covering brute force, DoS (both high-rate flooding and slow-rate), malformed packet injections, and MitM.
- We discuss existing tools that generate network flows from raw traffic and we discuss their pros and cons in this work's scope.
- We highlight our approach to generate a lightweight ML-based solution for malicious MQTT traffic detection.
- We explain limitations and challenges that affect the generation of high-quality ML-based tools.

The rest of this paper is organized as follows: we first provide a background on MQTT in Section II, then we review related work on MQTT security and intrusion detection (Section III). We follow that by describing aspects related to setup and data generation in Section IV; this includes attack types, dataset preparation, feature extraction, and ML techniques. Section V presents our approach, the experimental results, comparing performance across tools and addressing feature significance. Section VI provides a discussion of the findings, limitations, and future improvement avenues. Finally, Section VII concludes the paper and outlines future work.

II. MQTT Background

MQTT (Message Queuing Telemetry Transport) is a lightweight, publish-subscribe messaging protocol designed for efficient communication in environments with limited bandwidth, high latency, or unreliable connections. It is especially well-suited for resource-constrained devices such as sensors and actuators, making it a widely adopted protocol in IoT applications. MQTT is built to minimize network and device resource usage while still providing reliable message delivery.

The protocol operates on a client-broker architecture, where clients connect to a broker to either publish messages to specific topics or subscribe to topics to receive relevant data. Messages are organized into hierarchical topics, and the broker ensures that published messages are forwarded to all subscribed clients. This decoupled design enables scalable, bi-directional communication without direct device-to-device interaction. MQTT also supports retained messages and the Last Will and Testament (LWT) feature, which helps manage device state and unexpected disconnections.

MQTT provides three Quality of Service (QoS) levels—0 (at most once), 1 (at least once), and 2 (exactly once)—allowing developers to balance reliability and performance based on application needs. It supports secure communication through TLS/SSL encryption and offers authentication and authorization via credentials or certificates. Additionally, MQTT maintains continuous stateful sessions, remembering client subscriptions and undelivered messages, which is essential for devices with intermittent connectivity. These features collectively make MQTT a robust and flexible protocol for real-time, secure, and efficient IoT communication [8].

II-A. Attack Types

Like other network protocols, MQTT is susceptible to attacks. The literature discusses several attack types. Some of these attacks are MQTT specific, while others exist also in other areas. Table I demonstrates the types of attacks that exist in the MQTT context [9], [10], [7], [11].

Tabla I: Common MQTT Attack Types

Attack Type	**Description**
Brute-Force Attack	Attempts multiple username/password combinations to gain unauthorized access, especially when weak credentials are used.
DoS Attack	Floods the broker with excessive `CONNECT` or `PUBLISH` messages to exhaust system resources.
Slow DoS (SlowITe)	Maintains open connections by sending data slowly, gradually consuming broker resources until failure.
Malformed Packet Attack	Sends crafted or malformed MQTT packets to exploit parser or implementation vulnerabilities in the broker.
Man-in-the-Middle (MitM)	Intercepts unencrypted MQTT traffic to eavesdrop, inject, or alter messages in transit.
Port Scanning	Identifies MQTT brokers by probing standard ports (e.g., 1883), typically as a precursor to targeted attacks.
Topic Hijacking	Gains unauthorized access to publish or subscribe to topics, potentially leaking sensitive information or spoofing data.

III. Related Work

The literature includes several works that discuss different aspects of MQTT security and attacks. This includes vulnerabilities that have been found, security challenges and approaches to secure MQTT devices.

III-A. MQTT Security and Attacks

The security of MQTT in IoT has attracted increasing attention in recent years. MQTT operates at the application layer and is known to be suffer from vulnerabilities that require certain protection countermeasures. For instance, based on their analysis of MQTT-related vulnerabilities reported in the National Vulnerability Database (NVD) and the Common Vulnerabilities and Exposures (CVE), Husnain et al. [12] found 81 vulnerabilities between April, 2014 and December 2021. They classified these vulnerabilities into 3 main categories (described below) and 1 miscellaneous:

- *Required Fields Checks* (37 vulnerabilites): In MQTT, where required fields depend on the packet type, missing validations —such as a password field when a username is present— can cause serious issues. Multiple vulnerabilities demonstrate these risks, including server crashing, buffer overflows, and remote code execution.
- *Lack of Logical Errors Checks* (25): Absence of logical error checks in MQTT implementations can have severe impacts like crashes, buffer overflows, and DoS attacks when processing crafted packets. The reported vulnerabilities demonstrate how improper validation of fields like *Keep-Alive* and *Will Delay Interval* can be exploited.
- *Improper Packet Length Checks* (12): Packet parsing extracts and identifies packet fields, but without proper length checks, sequential parsing can be exploited by attackers using crafted packets. These reported vulnerabilities can lead to DoS, remote code execution, buffer overflow attacks, and unauthorized access to memory.

Thus, Husnain et al. [12] propose an MQTT parsing engine that can integrate with MQTT network-based IDS and maybe used for extensive checking against the traffic by implementing rigorous validation of packet fields when packets are parsed. They propose an engine that with 5 key steps: *a) protocol intuition, b) vulnerability assessment, c) protocol identification, d) parsing*, and *e) strict validation*.

Similarly, Hintaw et al. (2019) [5] studied MQTT's security challenges and solutions. They analyzed security considerations at different layers of the IoT architecture using MQTT:

- *Physical Layer*: running systems on private networks is an inherent security solution for industrial usage. Suggestions were made to use peer-to-peer structured networks and point-to-point networks to limit shared communication channels and potential hazards. Such measures might work by segregating physical units but are not practical when sensors are in unrestricted areas.
- *Network Layer*: IPSec is adequate for authentication and confidentiality but it can cause overloading due to extra headers. This may be mitigated by implementing 6LowPAN compression and using strong cryptographic schemes. Another option is the Host Identity Protocol (HIP) where broker and node identifiers can be separated from their locators. Yet, it might be challenging to employ cryptographic schemes and public key distribution.
- *Transport Layer*: Due to extra workload for setting up secure connections and encrypting all traffic, implementing TLS on resource-constrained MQTT devices is often not possible. In such cases, TLS Session Resumption can be used to significantly reduce the repetition of the TLS Handshake process. Also, computational load of ciphering can be reduced by using hardware-based TLS implementation, but this increases device cost and power consumption. OASIS, the entity in charge of the MQTT standard, recommends using TLS and certificates whenever possible to ensure confidentiality, integrity, and authentication [3]. The authors note that even though using TLS is a means of protection of MQTT traffic, improper or invalid TLS configuration can significantly lower security. They highlight that traffic might be compromised by exploiting TLS when known vulnerable TLS versions are used for the seek of maintaining compatibility with legacy devices within the network.
- *Application Layer*: An external authentication system

like *OAuth* or *LDAP* can be used for brokers in systems with multiple connecting items. In these scenarios, the external system is responsible for allowing access to tokens and authentication based on provided credentials. However, implementing such systems introduces extra complexity, and the pros and cons must be carefully considered during the design phase. Disadvantages may also include added latency and misconfiguration.

All in all, the authors divide the drawbacks of current security mechanisms in 3 categories. First, *High Complexity* is expected from the use of the different suggested security mechanisms, which is also reflected in the execution of processes and resource consumption. More complex mechanisms require more resources for operations like MQTT payload encryption-decryption. Second, only *Partial Protection for MQTT Protocol* exists. While some mechanisms address complexity issues, they might introduce new vulnerabilities. For instance, one security protocol was designed for having a fixed master key imprinted in devices. This protocol uses a challenge-based shuffling algorithm at both the client and server ends to make the key dynamic. This approach proved to be vulnerable. Even more, the dynamic session keys might remain exposed for a while, allowing attackers to decipher messages within the network. Third, *Additional Service Requirements* arise as the data volume grows. Traditional encryption schemes that encrypt MQTT payloads at the application layer are viable when devices have sufficient resources, but these schemes are limited in the areas of key management scalability, fine-grain access control and flexible data sharing.

III-B. Intrusion and Malicious Traffic Detection Systems for MQTT

Traditional IDS techniques often struggle with IoT protocols like MQTT, due to their focus on enterprise traffic patterns and the limited visibility into IoT device behavior. This is reflected in the fact that MQTT benign and malicious traffic are hard to differentiate using traditional IDS [7]. In response, specialized IDS systems for MQTT networks were proposed. Hindy et al. used their own developed MQTT traffic dataset, which features different MQTT attacks, to propose their specialized IDS. The authors tested multiple ML models on their dataset. Their results showed that their techniques can achieve higher detection rates for MQTT attacks than generic IDS. These improved results emphasize the importance of flow-based features, e.g. packet interarrival time, over simple packet-level features for distinguishing MQTT attacks from normal traffic.

Other works have explored traditional ML as well as Artificial Intelligence (AI) techniques for MQTT intrusion detection. Alzahrani and Aldhyani [13] evaluated several ML algorithms: *K-Nearest Neighbor (KNN)* and *Latent Dirichlet Allocation (LDA)* and deep learning models: *Convolutional Neural Network (CNN)* and *CNN Long Short-Term Memory Networks (CNN-LSTM)* on an MQTT attacks dataset. Their results show that the CNN-LSTM model had the highest accuracy (98 %) in identifying MQTT attacks as compared to the other models, which were 76 % (LDA), 80.3 % (CNN) and 80.8 % (KNN). This result suggests that complex nonlinear models can capture subtle patterns in MQTT traffic. However, deep models come at the cost of higher computational requirements, which might be impractical for deployment on IoT edge devices or constrained brokers.

Another study by Sultan et al. [14] specifically addressed MitM attack detection in MQTT networks using five different ML algorithms: *Aritificial Neural Network (ANN), Naïve Bayes (NB), Decision Trees (DT), KNN* and *eXtreme Gradient Boosting (XGBoost)*. Accuracy results ranged between 95.9 % for KNN and 99.6 % for XGBoost. The training time ranged between 0.008 seconds for KNN and 8.49 seconds for ANN. Thus, it reflects that while it is important to secure the underlying network, sometimes trade-offs have to be made for having a good performance, particularly within resources-constraint environment like MQTT Networks.

IV. Setup and Data Generation

In order to build data-driven platforms or tools that detect attacks on an MQTT network e.g. IDS, it is crucial to have an adequate setup (real-world or simulated) that generate data that represent real-world possible attacks. In this section, we discuss existing tools used in generating MQTT traffic (both benign and malicious) as well as existing datasets that were generated using such tools.

IV-A. Environment

The literature shows different approaches with regards to the environment used to generate data. The testbeds can either be fully real-world, fully simulated or a mix of both. GothX, for instance, is a testbed that includes a network of 450 emulated sensors [15]. On the other hand, CICIoMT2024 was built using a mix of emulated and real-world Internet of Medical Things devices [10].

IV-B. Traffic Generation Tools

The datasets MQTTset and CICIoMT2024 [16], [10] mention the use of IoT-Flock [17] in their MQTT traffic generation. Also, other tools are used to generate malicious traffic and they include:

- *MQTT-Malaria*: flooding DoS traffic
- *MQTTSA*: malformed data and brute-force authentication
- *MQTT-PWN*: brute-force authentication

IV-C. Datasets

While there are a lot of datasets that are provided in the literature, these three datasets stand out for their emphasis on MQTT network traffic (benign and malicious):

- *MQTTset* [16]: This dataset was created to address the limited availability of datasets suitable for IoT environments, particularly those supporting dedicated IoT protocols like MQTT. MQTTset was built using IoT-Flock, a network traffic generator tool capable of emulating IoT devices and networks based on MQTT and CoAP protocols. IoT-Flock also simulates various attacks against MQTT and CoAP, which were used to generate malicious traffic in MQTTset. The dataset simulates a smart home/office/building environment. The network consists of 8 different IoT sensors connected to an MQTT broker. These sensors are uniquely identified by IP addresses, and they retrieve information such as temperature, light intensity, among others. MQTTset includes both legitimate and malicious traffic. Besides

the normal traffic packet captures (pcaps), it includes traffic for the following attack types:

- **Flooding DoS**
- **Publish Flood**
- **SlowITe**
- **Malformed Data**
- **Brute-Force Authentication**

The captured traffic spans a temporal window of one week and consists of around 12M network packets with an overall size of about 1 GB. This dataset was reproduced by Poisson et al. [15] in their open-access GothX testbed.

- *MQTT-IoT-IDS2020* [7]: Introduced by Hindy et al., it is the first dataset to include MQTT-specific benign and attack instances. It simulates an IoT network that includes 12 MQTT sensors, a broker, a camera simulator, and an attacker. The four attack types are:
  - **Aggressive scan (Scan A)**: UDP scan using Nmap.
  - **UDP scan (Scan sU)**: Another UDP scan with Nmap.
  - **Sparta SSH brute-force (Sparta)**: Brute-force attack on SSH services.
  - **MQTT brute-force (MQTT BF)**: Targeting MQTT, performed with MQTT-PWN.
- *CICIoMT2024* [10]: This dataset is a benchmark for multi-protocol security assessment in the Internet of Medical Things (IoMT). It captures network traffic across Wi-Fi, MQTT, and Bluetooth, providing a diverse dataset for cybersecurity research. A significant portion of the dataset includes MQTT-based network activity, making it relevant for MQTT security research. Also, this dataset includes profiling traffic. This refers to traffic captured by extensive profiling of device states (idle, active, interaction). The testbed includes 40 IoMT devices (25 real, 15 simulated), including sensors, monitors, and controllers. MQTT devices were simulated within the IoMT testbed. The dataset contains the following MQTT-targeted attacks:
  - **MQTT Connect Flood**: Overwhelms the broker with excessive connection requests.
  - **MQTT Publish Flood**: Saturates the broker with high-frequency publish messages.
  - **MQTT Malformed Data**: Injects malformed packets to exploit MQTT protocol vulnerabilities.

V. Traffic Analysis Approach

There are several approaches and tools to analyze traffic in the context of MQTT. Hindy et al. [7] indicated that the performance of the ML models was better when flow-level features were used instead of packet-level features. Thus, we highlight several tools that generate traffic flows which could then be later used to create ML models of malicious traffic detection:

- *CICFlowMeter*: A widely used open-source flow generator from CIC (Canadian Institute for Cybersecurity). CICFlowMeter reads pcap files and produces bidirectional flow records along with a rich set of statistical features [18]. For each flow, it computes over 80 features such as duration, total packets and bytes in forward (client-to-broker) and backward (broker-to-client) directions, packet length statistics (min, max, mean, standard deviation), inter-arrival times, flow bytes/s rates, and some TCP-specific flags counts, among others. These features have been used in many prior IDS studies (including the creation of CIC-IDS2017 [19]). We used textitNTLFlowyLyzer [20], a newer python-based version of CICFlowMeter enriched with 348 features, to extract the standard feature set for all flows in our datasets. It provides a wider feature set for training the ML models and has less incompatibility issues upon installation.
- *Argus* [21]: Argus (Audit Record Generation and Usage System) is a network flow monitoring tool that operates in real-time or offline on packet traces. Argus generates a summarized record for each network connection, with emphasis on timing and performance metrics. From Argus, it is possible to obtain features such as average packet size, duration, packets per second, bytes per second, as well as network performance metrics like jitter and loss (where applicable), and basic flags indicating session termination conditions. Argus data format is flexible, and the tool is known for its reliability and is used in many operational networks for traffic analysis [21]. In our analysis, Argus provided a lighter weight feature set compared to CICFlowMeter (fewer features, mostly focusing on throughput metrics and counts).
- *Zeek* [22]: Zeek (formerly known as Bro) is an open-source network security monitor that can be used to analyze traffic and log protocol activities [22]. Zeek is not a flow exporter per se, but it produces connection logs where each entry corresponds to a network flow (connection) with various attributes. For pcap datasets, Zeek can generate the `conn.log`, which yields features including connection duration, total bytes and packets in each direction, the connection state (e.g., S0, SF for TCP states), and basic TCP flags.
- *NetFlow (nfdump)* [23]: We use the term "NetFlow"to indicate the extraction of classical NetFlow features. NetFlow is originally a Cisco protocol for exporting flow information from routers. The NetFlow feature set typically includes 5-tuple identification and counters like total bytes, packets, start and end times. It is much more minimal compared to CICFlowMeter. For example, NetFlow v5 records have just a handful of fields (duration, bytes, packets, source/dest ports, etc.). A feature set analogous to Cisco's NetFlow can be obtained: flow duration, total bytes, total packets, and flags for TCP. This represents a very lightweight feature approach. Using of NetFlow-like features can help evaluate how well simple features perform versus more extensive features. If a model can achieve good detection with only NetFlow features, that suggests even resource-limited devices (which often support NetFlow) could be leveraged for MQTT attack detection.

V-A. Our Approach

In our work, we aim to create a tool that trains on existing malicious traffic in order to create a behavioral fingerprint of existing MQTT-specific attacks. To reach this objective,

we are exploring existing datasets that offer traffic traces of attacks on the MQTT-protocol exclusively. However, this presents a challenge for us given the scarcity of datasets that have MQTT attack traffic. While we pointed out several datasets in Section IV-C that include attacks on an MQTT network, not all the attacks represented in these datasets are attacks on the MQTT protocol. Some attacks are done on other lower layers like MQTT-IoT-IDS2020's UDP scan attack. Another concern is that even when available, datasets can be very unbalanced like in MQTTset's case where the normal traffic represents 94.7 % while SlowITe attack represents 4.6 %. Collectively, both have 99.3 % of the dataset's packets. Generally, it is expected that malicious traffic be a small portion of traffic in a network. However, when the size of the traffic to be studied is limited in the dataset, it becomes challenging to generate a realistic fingerprint based on low number of samples. Thus, finding or even generating traffic that belong to other types of attacks becomes crucial to be able to train our model using sufficient data. We consider the attack types in MQTTset as a starting point to create our tool. These types we are as follows:

- Brute-force Authentication Attacks
- DoS Attacks (both CONNECT AND PUBLISH)
- Malformed Packet Injection
- Slow DoS (SlowITe Attack)

Our pipeline includes dataset collection, followed by analyzing and aggregating together attacks that are similar in their nature to existing attacks on MQTT even if they are not MQTT-based. The next steps involves generating traffic flows as we discuss in the next section (Section V-B) and using feature ranking algorithms to identify most significant features. Based on that, we may generate behavioral fingerprints for attacks. To ensure the correctness and high-quality of our model, it will be tested on similar traffic traces on which it has not been trained before. A visual representation of the pipeline is provided in Figure 1.

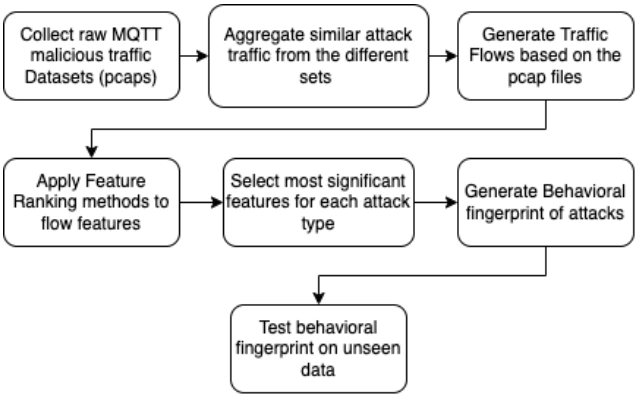


Figura 1: Pipeline of Behavioral Fingerprint of MQTT Attacks

V-B. Flow Generation

We are currently using two tools to generate our flows: Argus and NTLFlowLyzer. While both have some common fields, they are essentially different tools that have distinct sets of features. We note that while the performance of Argus is better than NTLFlowLyzer for some attacks, the contrary is also correct in some different situations. In this context, we note that while Argus takes minutes in processing large pcap files NTLFlowLyzer may take days to process a file of the same size. For example, in one case, a pcap of around 1 GB was processed in less than 10 minutes in Argus, while it took 7 days to generate the flow using NTLFlowLyzer.

V-C. Feature Ranking

Our approach is inspired by Hindy et al.'s [7] approach of identifying attacks using ML on MQTT traffic although it differs in the method and the ultimate goal. For instance, Hindy et al. use fields from the pcaps as features directly, while we convert pcaps into traffic flow format. Also, Hindy et al. use these features to test several ML algorithms. This is different from our goal, which is to identify the most significant features in order to build a lightweight classifier model that is capable of flagging attacks with the lowest possible cost while maintaining high accuracy.

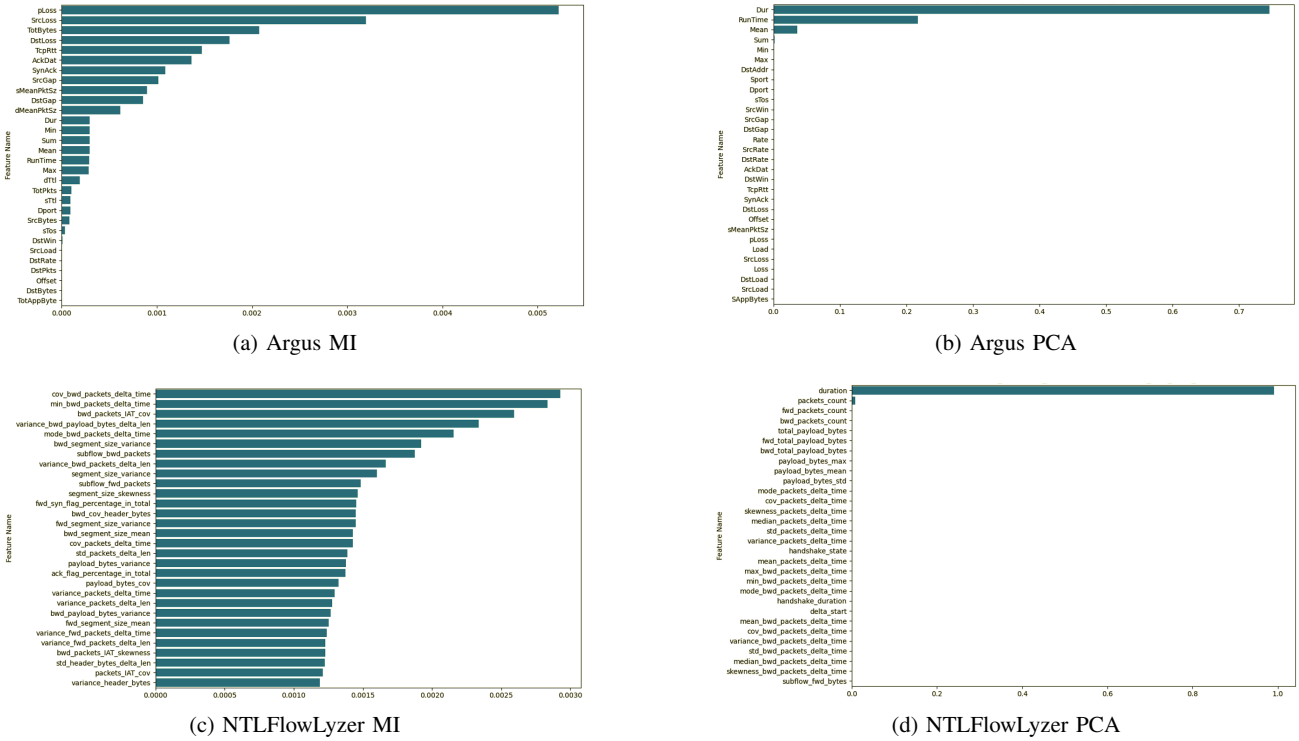
In order to be able to create a behavioral fingerprint for the MQTT-specific attacks, we seek to identify the most significant features in each type of attack. Thus, our approach focuses on testing different feature ranking algorithms. So far, we have tested 3 algorithms:

- *Mutual Information (MI)*: An algorithm that is used for feature selection.
- *Principal Component Analysis (PCA)*: An algorithm used for dimensionality reduction. It has been used in the literature for feature selection as well [24].
- *Laplacian Score*: An algorithm that is used for feature selection.

Laplacian Score was discarded for its very low performance. The preliminary results show that not always a specific algorithm is significant for all attacks. In fact, algorithms perform differently when tested with different types of attacks. MI may perform better or show higher influence of given features for one type of attack, while PCA may perform better for other types of attacks. Figure 2 shows a comparison of features in the BruteForce attack scenario from MQTT-IoT-IDS2020. It shows difference among the tools for the same algorithm and among the algorithms for the same tool. Even though *Duration* as a feature is very significant in both the PCA figures, *Runtime* and *Mean* have non-negligible weights in the Argus case (Figure 2b) compared to NTLFlowLyzer PCA (Figure 2d). We have carried out similar tests on different attack types in other datasets and highlight that outputs from the tools and the algorithm tend to vary based on the attack type.

VI. Discussion

As seen in the related work, two methodologies for malicious traffic detection are shown: 1) anomaly detection (learning normal MQTT traffic patterns and detecting deviations) and, 2) misuse detection (training classifiers to recognize known attack signatures). The majority of recent works, including those mentioned above, follow the latter approach by using labeled datasets to train supervised models. This requires that representative attack data are available for training. One limitation here is the availability of public datasets. Until 2020, there were no publicly available MQTT-specific intrusion datasets, which hindered research [7]. The



(a) Argus MI

(b) Argus PCA

(c) NTLFlowLyzer MI

(d) NTLFlowLyzer PCA

Figura 2: MQTT-IoT-IDS2020 Brute-Force Attack Feature Ranking using Argus and NTLFlowLyzer using MI and PCA

creation of datasets like MQTT-IoT-IDS2020 [7] and MQTT-Tset [16] has filled this gap, but each has a limited set of attack scenarios. Generalization of models beyond the training dataset is a concern – for example, a model trained on a synthetic smart home MQTT dataset might not directly be fit for an industrial IoT MQTT deployment. Hindy et al. noted that MQTT-based attacks can mimic normal operations closely, making generalization challenging [7].

Another limitation in existing approaches is the computational overhead. Some studies using deep learning (e.g., CNN, LSTM) assume the availability of substantial processing power, which might not be feasible on actual IoT network nodes. There is often a trade-off between detection accuracy and complexity. Simpler models (KNN, decision trees) are faster but might yield more false negatives, whereas complex ensembles or neural networks perform better at the cost of speed and transparency. Real-time detection is not addressed in many papers; models are evaluated offline on captured data. For deployment, an IDS must process MQTT traffic on the fly with minimal latency, which remains an open challenge. Moreover, existing research typically focuses on a single feature extraction method (for instance, using only CICFlowMeter features). There is limited exploration into how different feature sets or tools might impact detection efficacy. This is one aspect we address in our work by studying multiple flow analysis tools.

Overall, related work shows that ML is a promising approach for detecting MQTT network attacks, achieving high accuracies on known datasets. Nonetheless, gaps remain in terms of dataset diversity, MQTT-specific attacks, model generalizability, and practical deployment. Our work builds on this foundation by examining multiple datasets, feature extraction techniques, and algorithms side-by-side, and by discussing the feasibility of integrating such a detection system into real MQTT environments. Our preliminary results indicate differences between analysis tools and indicate variability when using them. Knowing that attacks, even unseen ones, have some features that relate to previously seen ones, our goal is to build a model attacks on which it was not trained before.

If successful, our approach can be the basis for a lightweight detection framework that has acceptable performance in terms of accuracy. In this sense, building a solution that relies on misuse detection as an approach is a more realistic one that fits the limited-resource networks in which MQTT is typically implemented and that could be an adequate protection solution for IoT edge networks which tend to be limited in processing and in power resources.

VII. Conclusion

In this work, we aimed to shed the light on MQTT's security and challenges arising from developing protection solutions against attacks on the protocol. To this end, we discussed the protocol and the attack vectors (generic and protocol-specific). Also, we addressed setup for generating normal and malicious traffic, existing datasets, and tools for flow generation. Finally, we explained our planned approach to create a tool that is capable of identifying malicious traffic by pinpointing the most significant features in each attack type using feature ranking.

Acknoledgements

This work was supported by the grant COMPROMISE (PID2020-113795RB-C32) and funded

by MICIU/AEI/10.13039/501100011033; the grant DISCOVERY (PID2023-148716OB-C33) funded by MCIU/AEI/10.13039/501100011033 and FEDER, UE and the Grant QURSA (TED2021-130369B) funded by MICIU/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR. This research results from the I-Shaper Strategic Project (C114/23), due to the collaboration agreement signed between the Instituto Nacional de Ciberseguridad (INCIBE) and the Universidad Carlos III de Madrid. It is worth highlighting that this initiative is being carried out within the framework of the Recovery, Transformation and Resilience Plan funds, funded by the European Union (Next Generation).

Referencias

[1] "Internet of things: Key stats for 2022," accessed: 21 March 2025. [Online]. Available: https://techinformed.com/internet-of-things-key-stats-for-2022/#:~:text=According%20to%20Statista%2C%20the%20number,44%20billion%20by%202025

[2] "What is mqtt?" accessed: 20 March 2025. [Online]. Available: https://aws.amazon.com/what-is/mqtt/?nc1=h_ls

[3] "Mqtt version 5.0," accessed: 20 March 2025. [Online]. Available: https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.html

[4] "8 real-world mqtt use cases," accessed: 19 March 2025. [Online]. Available: https://www.influxdata.com/blog/mqtt-use-cases/

[5] A. J. Hintaw, S. Manickam, S. Karuppayah, and M. F. Aboalmaaly, "A brief review on mqtt's security issues within the internet of things (iot)." *J. Commun.*, vol. 14, no. 6, pp. 463–469, 2019.

[6] H. Wong and T. Luo, "Man-in-the-middle attacks on mqtt-based iot using bert based adversarial message generation," in *KDD 2020 AIoT Workshop*, vol. 8, 2020.

[7] H. Hindy, E. Bayne, M. Bures, R. Atkinson, C. Tachtatzis, and X. Bellekens, "Machine learning based iot intrusion detection system: An mqtt case study (mqtt-iot-ids2020 dataset)," in *International networking conference*. Springer, 2020, pp. 73–84.

[8] "Mastering mqtt: The ultimate beginner's guide for 2025," accessed: 19 March 2025. [Online]. Available: https://www.emqx.com/en/blog/the-easiest-guide-to-getting-started-with-mqtt#what-is-mqtt

[9] "Reinforcing security of ot systems in iiot with mqtt and hivemq," accessed: 18 March 2025. [Online]. Available: https://www.hivemq.com/blog/reinforcing-security-ot-systems-iiot-mqtt-hivemq/

[10] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. Ghorbani, "Ciciomt2024: Attack vectors in healthcare devices-a multi-protocol dataset for assessing iomt device security," 2024.

[11] I. Vaccari, M. Aiello, and E. Cambiaso, "Slowite, a novel denial of service attack affecting mqtt," *Sensors*, vol. 20, no. 10, p. 2932, 2020.

[12] M. Husnain, K. Hayat, E. Cambiaso, U. U. Fayyaz, M. Mongelli, H. Akram, S. Ghazanfar Abbas, and G. A. Shah, "Preventing mqtt vulnerabilities using iot-enabled intrusion detection system," *Sensors*, vol. 22, no. 2, p. 567, 2022.

[13] A. Alzahrani and T. H. Aldhyani, "Artificial intelligence algorithms for detecting and classifying mqtt protocol internet of things attacks," *Electronics*, vol. 11, no. 22, p. 3837, 2022.

[14] A. B. M. Sultan, S. Mehmood, and H. Zahid, "Man in the middle attack detection for mqtt based iot devices using different machine learning algorithms," in *2022 2nd International Conference on Artificial Intelligence (ICAI)*. IEEE, 2022, pp. 118–121.

[15] M. Poisson, R. Carnier, and K. Fukuda, "Gothx: a generator of customizable, legitimate and malicious iot network traffic," in *Proceedings of the 17th Cyber Security Experimentation and Test Workshop*, 2024, pp. 65–73.

[16] I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso, "Mqttset, a new dataset for machine learning techniques on mqtt," *Sensors*, vol. 20, no. 22, p. 6578, 2020.

[17] S. Ghazanfar, F. Hussain, A. U. Rehman, U. U. Fayyaz, F. Shahzad, and G. A. Shah, "Iot-flock: An open-source framework for iot traffic generation," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*. IEEE, 2020, pp. 1–6.

[18] A. H. Lashkari, A. Seo, G. D. Gil, and A. Ghorbani, "Cic-ab: Online ad blocker for browsers," in *2017 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2017, pp. 1–7.

[19] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani *et al.*, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." *ICISSp*, vol. 1, no. 2018, pp. 108–116, 2018.

[20] M. Shafi, A. H. Lashkari, and A. H. Roudsari, "Ntlflowlyzer: Towards generating an intrusion detection dataset and intruders behavior profiling through network and transport layers traffic analysis and pattern extraction," *Computers & Security*, vol. 148, p. 104160, 2025.

[21] "argus," accessed: 21 March 2025. [Online]. Available: https://openargus.org/

[22] "The zeek network security monitor," accessed: 21 March 2025. [Online]. Available: https://zeek.org/

[23] "nfdump," accessed: 13 May 2025. [Online]. Available: https://github.com/phaag/nfdump

[24] C. Pascoal, M. R. De Oliveira, R. Valadas, P. Filzmoser, P. Salvador, and A. Pacheco, "Robust feature selection and robust pca for internet traffic anomaly detection," in *2012 Proceedings Ieee Infocom*. IEEE, 2012, pp. 1755–1763.

Análisis y Gestión de riesgos de ciberseguridad en entornos IoT: El patrón MARISMA-IoT

José Luis Ruiz Catalán, Ferney Martinez, Ángel Suárez-Bárcena,
Antonio Santos-Olmo, David G. Rosado, Luis E. Sánchez Crespo
GSyA Research Group. Universidad de Castilla-La Mancha, Ciudad Real, España
joseluis.ruiz@uclm.es, ferney.martinez@alu.uclm.es,
{angel.sbarcena, antonio.santosolmo david.grosado, luise.sanchez}@uclm.es

***Resumen*—Los sistemas IoT optimizan la eficiencia, seguridad y gestión de datos, pero su expansión implica desafíos de ciberseguridad, ya que una vulnerabilidad puede comprometer toda la infraestructura. Las metodologías tradicionales de evaluación de riesgos no se ajustan a la naturaleza distribuida y dinámica de estos ecosistemas. Este trabajo propone un enfoque basado en MARISMA y eMARISMA, su entorno en la nube, para gestionar riesgos en IoT. Definimos un patrón de riesgos alineado con marcos como NIST, ENISA e ISO 27001 y presentamos un caso de estudio en un hogar inteligente. Nuestra solución ofrece una metodología flexible y escalable para mitigar riesgos en IoT, garantizando resiliencia ante amenazas y facilitando la integración de nuevos dispositivos.**

***Index Terms*—Ciberseguridad, Análisis y gestión de riesgo, IoT, Smart Home, MARISMA, NIST, ISO, ENISA.**

Tipo de contribución: Investigación original

I. Introducción

Los sistemas IoT han transformado la forma en que operan las infraestructuras conectadas, optimizando la eficiencia, la sostenibilidad y la seguridad en diversos entornos. Estas redes inteligentes incluyen sensores, dispositivos de automatización, sistemas de control de acceso, gestión energética y otras tecnologías avanzadas que permiten optimizar el uso de recursos y mejorar la experiencia del usuario [1][2]. Sin embargo, la creciente interconectividad de estos dispositivos ha incrementado las vulnerabilidades de seguridad [3][4], convirtiendo a las infraestructuras IoT en objetivos cada vez más atractivos para los ciberataques.

Los métodos tradicionales de evaluación de riesgos en TI no pueden aplicarse directamente a los entornos IoT debido a la combinación de elementos físicos y digitales, la diversidad de protocolos y dispositivos, y la necesidad de garantizar alta disponibilidad y fiabilidad en sus operaciones. En este contexto, la gestión de riesgos debe abordar no solo la protección de datos, sino también la seguridad física y la integridad operativa de los sistemas IoT [5] [6].

Para enfrentar estos desafíos, proponemos un marco de gestión de riesgos basado en MARISMA (Metodología para el Analisis de Riesgos en Sistemas de Información, utilizando Meta-Patron y Adaptabilidad) [7], una metodología que permite identificar, evaluar y mitigar los riesgos de ciberseguridad en entornos dinámicos y complejos. Una de las ventajas de MARISMA es su capacidad de adaptación a diferentes entornos, como se puede encontrar en trabajos previos donde MARISMA se adaptó a un entorno de Big Data [8], a un entorno de CPS [9] y a procesos de negocio [10], aplicándose exitosamente a casos reales.

A través del desarrollo del patrón MARISMA-IoT, nuestro enfoque proporciona un modelo adaptable y reutilizable para evaluar amenazas, activos y vulnerabilidades dentro del ecosistema IoT. Además, alineamos nuestra propuesta con regulaciones y estándares clave, como ISO/IEC 27001:2023 [11], ISO/IEC 27005:2022 [12], la Baseline Security Recommendations for IoT in the context of Critical Information Infrastructures (ENISA, 2017) [13], la ENISA Threat Landscape and Good Practice Guide for Smart Home and Converged Media [14] y el marco de NIST para IoT (NIST 8259 (Foundational Cybersecurity Activities for IoT Device Manufacturers) [15], NIST 800-82 (Guide to Industrial Control Systems Security) [16] y NIST 800-161 (Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations) [17]).

Esta investigación introduce el patrón MARISMA-IoT y lo valida mediante un caso de estudio aplicado a un entorno doméstico inteligente, en el que se demuestra su utilidad en la identificación y mitigación de riesgos específicos de los sistemas IoT. Los resultados obtenidos muestran que el enfoque propuesto no solo facilita la evaluación de riesgos en tiempo real, sino que también optimiza la toma de decisiones en la gestión de la seguridad en infraestructuras IoT.

En la elaboración del patrón MARISMA-IoT se ha revisado el patrón MARISMA-CPS [9] ya que comparten ciertos puntos en común, debido a que ambos consideran la integración de dispositivos IoT, pero con enfoques claramente diferenciados. MARISMA-IoT se especializa en entornos donde los dispositivos IoT son el elemento central y principal a proteger, mientras que MARISMA-CPS se orienta hacia infraestructuras más amplias donde los dispositivos IoT son solo una parte del ecosistema. Otra diferencia clave radica en los estándares utilizados para la definición de sus componentes y taxonomías. MARISMA-IoT adopta el estándar ISO/IEC 27001:2023, ampliamente reconocido en el ámbito de la ciberseguridad, para definir sus dominios y objetivos de protección. En contraste, MARISMA-CPS se basa en las recomendaciones de ENISA 2017 (Baseline Security Recommendations for IoT in the context of Critical Information Infrastructures), un marco de referencia centrado en la seguridad de infraestructuras críticas que incluyen IoT, pero sin abordar en profundidad todas sus dimensiones y amenazas específicas. Además, la cobertura de riesgos en cada patrón es distinta. MARISMA-CPS contempla los riesgos derivados de la integración de dispositivos IoT, pero su alcance no es completo, lo que significa que no aborda todas las posibles amenazas ni las múltiples dimensiones que pueden afectar a

una infraestructura con IoT como elemento clave. Por lo tanto, MARISMA-IoT se diseña específicamente para proporcionar una gestión de riesgos integral en entornos donde los dispositivos IoT son fundamentales, garantizando un enfoque más detallado y adaptado a sus particularidades.

Este artículo se estructura de la siguiente manera: En la Sección II, se presenta una introducción con los antecedentes del tema y las referencias utilizadas para la definición del patrón. En la Sección III, se describe el marco de trabajo de MARISMA, destacando sus principales características y el proceso para definir el patrón. En la Sección IV, se detalla el patrón MARISMA-IoT, explicando cada uno de sus elementos. En la Sección V, se expone un caso de estudio y se analizan los resultados obtenidos tras la aplicación de MARISMA-IoT. Finalmente, en la Sección VI, se presentan las conclusiones y las posibles líneas de trabajo futuro.

II. Trabajo relacionado

La creciente adopción de sistemas IoT ha impulsado la necesidad de marcos de gestión de riesgos específicos que aborden la ciberseguridad en entornos altamente interconectados. Diversos estándares y regulaciones han sido desarrollados para mejorar la seguridad y confiabilidad de estos sistemas, proporcionando directrices para la protección de infraestructuras críticas y dispositivos IoT. A continuación, presentamos una revisión de los principales estándares y marcos regulatorios utilizados en la construcción de nuestro enfoque.

II-A. Estándares internacionales de seguridad y gestión de riesgos

Uno de los marcos de referencia más importantes en la gestión de seguridad de la información es ISO/IEC 27001:2023 [11], que establece los requisitos para implementar y mantener un Sistema de Gestión de Seguridad de la Información (SGSI). Esta norma proporciona un enfoque basado en el riesgo, asegurando la confidencialidad, integridad y disponibilidad de la información en entornos críticos que se puede adaptar a los entornos que integren dispositivos IoT.

Por su parte, ISO/IEC 27005:2022 [12] complementa este enfoque proporcionando directrices específicas para la evaluación de riesgos en infraestructuras inteligentes, garantizando la resiliencia operativa de los dispositivos conectados.

II-B. Marcos de ciberseguridad del NIST para IoT y entornos industriales

El NIST (National Institute of Standards and Technology) ha desarrollado múltiples marcos de referencia para abordar la seguridad en sistemas IoT y entornos industriales:

- NIST 8259 (Foundational Cybersecurity Activities for IoT Device Manufacturers) [15]: Proporciona un enfoque estandarizado para la seguridad de dispositivos IoT, estableciendo prácticas esenciales para la protección de estos sistemas frente a amenazas cibernéticas desde el punto de vista del fabricante.
- NIST 800-82 (Guide to Industrial Control Systems Security) [16]: Se centra en la seguridad de los sistemas de control industrial (ICS), proporcionando directrices clave para proteger infraestructuras críticas como los sistemas de gestión de edificios (BMS, Building Management Systems), los cuales pueden ser integrados en las infraestructuras que contengan dispositivos IoT con demandas de seguridad y de funcionamiento más críticos.
- NIST 800-161 (Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations) [17]: Aborda la seguridad en cadenas de suministro en cualquier sector, pudiendo esta ser aplicada en entornos IoT donde múltiples proveedores y fabricantes intervienen en la provisión de hardware y software. En el contexto de los Smart Buildings, este estándar ayuda a mitigar riesgos asociados con la integración de dispositivos de diferentes fabricantes y las presentes amenazas de la adquisición de los dispositivos sin la comprobación pertinente del proveedor.

II-C. Recomendaciones de seguridad para IoT

Además de los estándares mencionados, existen directrices específicas para la protección de dispositivos IoT en infraestructuras críticas. La Baseline Security Recommendations for IoT in the context of Critical Information Infrastructures (ENISA, 2017) [13] proporciona un conjunto de medidas de seguridad fundamentales para dispositivos conectados en entornos críticos. Estas recomendaciones son especialmente relevantes en entornos IoT, donde la conectividad de sistemas como control de accesos, HVAC, videovigilancia y sensores ambientales puede ser un vector de ataque.

Por otro lado, la ENISA Threat Landscape and Good Practice Guide for Smart Home and Converged Media [14] analiza las amenazas emergentes en entornos domésticos inteligentes y plataformas de medios convergentes, proporcionando buenas prácticas aplicables a los entornos domésticos inteligentes. Este documento destaca el impacto de amenazas como el secuestro de dispositivos IoT, ataques de denegación de servicio (DDoS) y la manipulación de datos en sensores críticos.

II-D. Enfoques previos para la gestión de riesgos en infraestructuras IoT

Diversos enfoques han sido propuestos para evaluar y mitigar los riesgos en infraestructuras inteligentes. Algunos estudios han adaptado metodologías de análisis de riesgos tradicionales, como MAGERIT, OCTAVE y CORAS, para entornos IoT. Sin embargo, estas metodologías presentan limitaciones en cuanto a adaptabilidad y automatización, dificultando su aplicación en ecosistemas IoT dinámicos.

En contraste, nuestro trabajo introduce un patrón de gestión de riesgos basado en MARISMA, que combina la evaluación estructurada de riesgos con la reutilización de conocimiento y la adaptabilidad a diferentes entornos IoT. A diferencia de otros enfoques, nuestra propuesta permite la gestión continua de riesgos, integrando herramientas automatizadas para la identificación y mitigación de amenazas en tiempo real.

III. MARISMA Framework

MARISMA [7] es un marco diseñado para la gestión dinámica de riesgos de ciberseguridad, enfocado en la adaptabilidad, el aprovechamiento del conocimiento previo y la reducción de la subjetividad en su evaluación. Su metodología integra un metapatrón de riesgos reutilizable junto con patrones específicos adaptados a distintos sectores, permitiendo un modelado flexible y escalable. El framework se basa en tres

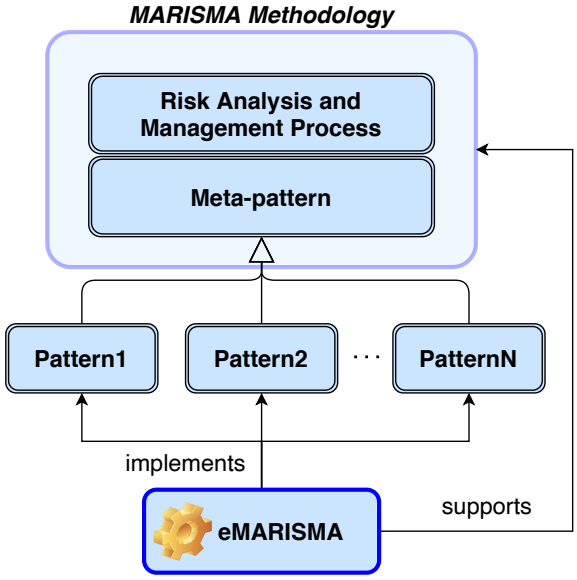


Figura 1. MARISMA Framework.

elementos principales: un metapatrón que define la estructura de controles, activos y amenazas; patrones específicos para distintos sectores; y procesos de gestión del riesgo que permiten desarrollar, aplicar y actualizar patrones en función de eventos en tiempo real. Además, MARISMA incorpora un enfoque de aprendizaje continuo, asegurando su adaptación constante a nuevos escenarios de riesgo. Para optimizar su aplicación, MARISMA cuenta con una herramienta en la nube (eMARISMA) que automatiza el análisis y la gestión de riesgos, ofreciendo accesibilidad, reducción de costos y tiempos de implementación, generación de alertas basadas en eventos de seguridad y escalabilidad para organizaciones de cualquier tamaño.

En la Figura 1 se puede observar como MARISMA integra un análisis y gestión de riesgos a traves del metra-patrón que establece la estructura de controles, el cuál es alimentado por patrones específicos que proporcionan los activos y amenazas permitiendo un modelo flexible y escalable.

IV. Definición del patrón para infraestructuras IoT en MARISMA-IoT

El desarrollo de un patrón de gestión de riesgos en infraestructuras IoT requiere una estructura bien definida que permita evaluar y mitigar amenazas específicas en entornos altamente interconectados. En este trabajo, hemos utilizado como referencia los estándares internacionales más relevantes, incluyendo ISO/IEC 27001:2023 [11], ISO 27005:2022 [12], NIST 8259 [15], NIST 800-82 [16], NIST 800-161 [17], y las recomendaciones de ENISA Baseline Security Recommendations [13] y ENISA Threat Landscape [14].

El patrón propuesto está diseñado para proporcionar un enfoque adaptable y escalable en la evaluación de riesgos, considerando activos, amenazas, controles de seguridad y dimensiones clave para entornos IoT, como se puede ver en la Figura 2.

A continuación, se describen los principales componentes del patrón MARISMA-IoT, permitiendo su integración en infraestructuras inteligentes como oficinas corporativas, edificios residenciales, centros comerciales, etc.

IV-A. Dominios y Objetivos

Para definir los dominios y objetivos de seguridad en sistemas que integren dispositivos IoT, hemos utilizado la ISO/IEC 27001:2023 como referencia principal. Esta norma establece un conjunto de controles de seguridad esenciales para garantizar la confidencialidad, integridad y disponibilidad de los activos digitales y físicos que hemos adaptado para un entorno IoT.

Cada dominio está asociado con objetivos de control, alineados con las recomendaciones de ISO/IEC 27001:2023, estableciendo controles específicos para gestionar los riesgos en edificios inteligentes tal y como se puede ver en la Figura 2.

IV-B. Activos

Los activos en un entorno IoT incluyen dispositivos físicos, infraestructura digital, plataformas de gestión y datos críticos, que pueden ser objeto de ataques o vulnerabilidades. Para la definición de los activos, hemos utilizado como referencia los documentos proporcionados por ENISA y pueden ser consultados en la Figura 2.

Estos activos son esenciales para el funcionamiento de una infraestructura IoT, como para ser catalogado como tal, y su protección es clave para garantizar la seguridad del ecosistema IoT.

IV-C. Dimensiones

La evaluación de riesgos en entornos IoT debe considerar múltiples dimensiones de seguridad, las cuales han sido definidas a partir de la ISO/IEC 27001:2023, el NIST 8259, 800-82 y 800-161. Estas dimensiones permiten un enfoque integral en la protección de infraestructuras inteligentes (dichas dimensiones se pueden observar en la Figura 2). Además, estas dimensiones aseguran que la evaluación de riesgos en sistemas IoT sea estructurada y efectiva, proporcionando un marco de referencia sólido para su protección.

IV-D. Amenazas

Las infraestructuras IoT están expuestas a una variedad de amenazas que pueden comprometer su seguridad y operación. Para definir estas amenazas, hemos tomado como referencia la ENISA Baseline Security Recommendations y ENISA Threat Landscape (dichas amenazas pueden ser consultadas en la Figura 2).

La presencia de estas amenazas requiere una estrategia de gestión de riesgos proactiva, que permita la identificación y mitigación de vulnerabilidades antes de que sean explotadas.

IV-E. Matriz de Dominios-Objetivos-Amenazas en infraestructuras IoT

La gestión de riesgos en infraestructuras IoT requiere una estructura clara que relacione los dominios de seguridad, los objetivos de control y las amenazas identificadas. Para ello, hemos construido una matriz de correlación que permite visualizar cómo cada amenaza impacta en diferentes dominios y objetivos de seguridad, como se muestra en la Tabla I.

Los dominios y objetivos han sido adaptados de la ISO/IEC 27001:2023, proporcionando un marco de referencia sólido

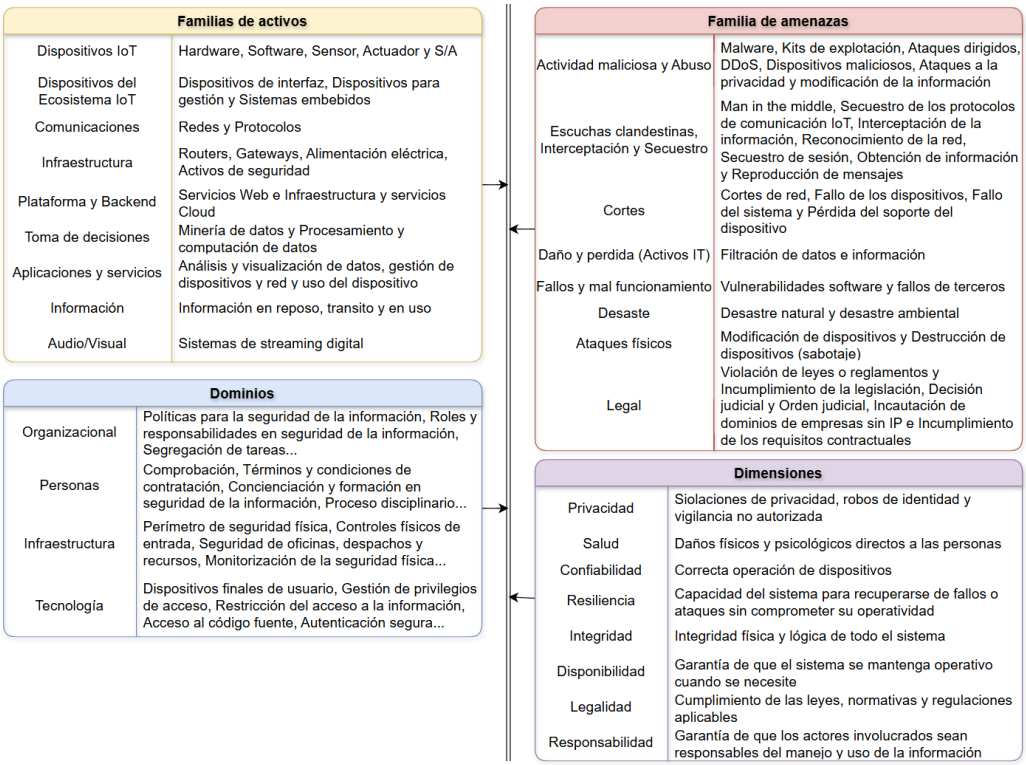


Figura 2. Elementos del patrón MARISMA-IoT.

para la protección de infraestructuras inteligentes. Por otro lado, las amenazas han sido extraídas de ENISA Baseline Security Recommendations y ENISA Threat Landscape, asegurando su alineación con los desafíos de seguridad específicos en entornos IoT.

Dado que la matriz completa es de gran extensión, en esta sección presentamos una versión reducida con ejemplos representativos en la Tabla I. Si nos fijamos en dicha tabla, podemos ver por ejemplo que, el dominio *Infraestructura* con el objetivo correspondiente a los *Controles físicos de entrada*, ayudan a mitigar los riesgos de *Daño / Pérdida IT, Ataques físicos y Legal*. Esto se debe a que, al controlar el acceso al espacio físico donde se encuentran los dispositivos IoT, se impide que personas no autorizadas puedan ingresar, evitando así posibles daños a los dispositivos o la extracción de información de ellos.

IV-F. Matriz de Activos-Amenazas-Dimensiones en infraestructuras IoT

La gestión de riesgos en entornos IoT requiere un análisis estructurado de los activos, amenazas y dimensiones de seguridad. Para ello, hemos construido una matriz que relaciona los tipos de activos, las amenazas que pueden afectarlos y las dimensiones de seguridad comprometidas. Para completar esta tarea, se ha utilizado el conocimiento y la amplia experiencia de los autores en ciberseguridad, y análisis y gestión de riesgos que fueron de gran utilidad para extraer información detallada sobre las relaciones entre la amenaza, los activos afectados y las dimensiones seguridad dentro de los dominios de seguridad definidos previamente.

La Tabla II muestra las conexiones entre los distintos tipos de amenazas y las dimensiones más vulnerables de cada tipo de activo. En la tabla se identifican tanto los activos como las amenazas que pueden afectarlos. Sin embargo, no todas las amenazas impactan a todos los activos. Es importante destacar que una amenaza tiene como objetivo principal perjudicar un activo, disminuyendo su valor. Dicho valor está determinado por sus dimensiones. Por ejemplo, si una amenaza afecta únicamente la privacidad de un activo, no dañará sus otras dimensiones. En otros casos, una amenaza puede comprometer la privacidad, la confiabilidad y la confidencialidad de un activo, sin afectar su integridad, disponibilidad o seguridad.

Dado que la matriz completa es de gran extensión, en esta sección presentamos una versión reducida con ejemplos representativos. Si observamos la tabla podemos ver que, por ejemplo, la familia de amenazas *Cortes* y el tipo de amenaza correspondiente a *Corte de red* afecta a las dimensiones *confiabilidad, resiliencia, disponibilidad, legalidad y responsabilidad* de los activos pertenecientes a las comunicaciones y los de infraestructura. Esto es debido a que un corte de red afectaría de la siguiente manera a la infraestructura teniendo en cuenta las dimensiones:

1. Confiabilidad (C): Puede afectar la estabilidad de los dispositivos de red y la conectividad del sistema.
2. Resiliencia (RL): Puede dificultar la recuperación del sistema si la interrupción persiste o se prolonga.
3. Disponibilidad (D): Puede hacer que la infraestructura de red quede inoperable, afectando todos los servicios conectados.
4. Legalidad (L): Puede incumplir regulaciones de disponibilidad si afecta servicios esenciales.
5. Responsabilidad (RP): Se requiere monitoreo continuo

Tabla I
SEGMENTO DE LA MATRIZ DOMINIOS, OBJETIVOS Y FAMILIAS DE AMENAZAS.

Dominio	Objetivo	Familias de Amenazas							
		Actividad Maliciosa	Escuchas / Interceptación	Cortes	Daño / Pérdida IT	Fallos / Mal funcionamiento	Desastre	Ataques físicos	Legal
Organizacional	Políticas para la seguridad de la información	X	X	X	X	X	X	X	X
	Roles y responsabilidades en seguridad de la información	X	X	X	X	X	X	X	X
	Segregación de tareas	X	X						X
Personas	Comprobación	X							X
	Términos y condiciones de contratación	X							X
	Concienciación, educación y formación en seguridad de la información	X	X		X	X			X
Infraestructura	Perímetro de seguridad física				X			X	X
	Controles físicos de entrada				X			X	X
	Seguridad de oficinas, despachos y recursos				X			X	X
Tecnología	Dispositivos finales de usuario	X	X		X				X
	Gestión de privilegios de acceso	X			X	X			X
	Restricción del acceso a la información	X	X		X	X			X

y planes de contingencia para asegurar la continuidad operativa.

V. CASO DE ESTUDIO: SMART HOME

Los Smart Homes han evolucionado hasta convertirse en entornos interconectados donde múltiples dispositivos IoT automatizan y optimizan tareas cotidianas, mejorando la eficiencia, la seguridad y la comodidad de los residentes. Sin embargo, la creciente interconectividad introduce nuevas vulnerabilidades de ciberseguridad, que pueden ser explotadas por atacantes para acceder a datos personales, manipular dispositivos críticos o comprometer la privacidad de los ocupantes.

Este caso de estudio aplica el patrón MARISMA-IoT para evaluar los riesgos y amenazas en un hogar inteligente con dispositivos IoT integrados, permitiendo identificar activos críticos, analizar posibles ataques y proponer controles de seguridad adaptados. Además se ha decidido que corresponda a un escenario donde el usuario o propietario de la Smart Home tiene bajos conocimientos informáticos y en su defecto, bajos conocimientos de ciberseguridad y securización de su infraestructura IoT, dando como resultado un escenario realista, ya que la mayoría de los usuarios instalan estos dispositivos en sus hogares, sin tener en cuenta ninguna medida de ciberseguridad, estando expuestos a multitud de riesgos.

V-A. Definiendo las dimensiones para el caso de estudio

Para evaluar los riesgos en un Smart Home, se han considerado todas las dimensiones de ciberseguridad definidas en el patrón ya que son fundamentales para evaluar el impacto de las distintas amenazas en los diferentes activos que componen el caso de estudio, los cuales se pueden ver en la Figura 2. Estas dimensiones permiten analizar la seguridad desde diferentes perspectivas.

V-B. Definiendo los activos para el caso de estudio

Un Smart Home incorpora una variedad de activos interconectados que gestionan seguridad, energía, automatización y comunicación. Muchos de los activos presentes en un Smart Home están presentes en un entorno doméstico tradicional



Figura 3. Segmento de los activos añadidos a la herramienta eMARISMA.

pero existen muchos otros que son específicos de este, que al estar conectados de forma inteligente, pueden tomar decisiones de forma autónoma. Los tipos de activos para el caso de estudio que se han considerado instanciandolos con los elementos previamente definidos en el patrón se pueden observar en la Tabla III.

Una vez identificados y clasificados, los activos se incorporan fácilmente a eMARISMA, que los organiza jerárquicamente en familias, tipos y activos definidos en el patrón, pudiendo ser observado en la Figura 3. La herramienta permite añadir información relevante, como el valor del activo, utilizado para priorizar y mejorar la precisión en la evaluación del riesgo.

Tras su incorporación, eMARISMA analiza los riesgos considerando las relaciones entre activos, amenazas y dimensiones (establecidas en la Tabla II), mostrando en tiempo real el nivel de riesgo y los controles recomendados para su protección.

V-C. Definiendo las amenazas para el caso de estudio

Para continuar con el análisis de riesgos, se identifican las amenazas que pueden afectar el ecosistema doméstico, basándose en el patrón MARISMA-IoT y en los activos a proteger (Tabla IV). La herramienta eMARISMA permite seleccionar estas amenazas, asignando valores de degradación del activo y probabilidad de ocurrencia (de 0 a 100) cómo se puede ver en la Figura 4 donde se ha establecido, por ejemplo, que los *ataques a la privacidad* tienen un porcentaje de ocurrencia del 60 %, ya que para que ocurra ese ataque debe de haber una serie de pasos anteriores que lo hacen

Tabla II
SEGMENTO DE LA MATRIZ DE FAMILIAS Y TIPOS DE AMENAZAS, FAMILIAS DE ACTIVOS, Y DIMENSIONES.

Familias de Amenazas	**Tipos de Amenazas**	**Familias de Activos**								
		Dispositivos IoT	Otros dispositivos IoT	Comunicaciones	Infraestructura	Plataforma y Backend	Toma de decisiones	Aplicaciones y Servicios	Información	Audio/Visual
Actividad maliciosa	Malware	P, S, C, RL, I, D, L, RP	P, S, C, RL, I, D, L, RP			P, S, C, RL, I, D, L, RP				P, S, C, RL, I, D, L, RP
	Kits de explotación	P, S, C, RL, I, D, L, RP	P, S, C, RL, I, D, L, RP		S, C, RL, I, D, L, RP					P, S, C, RL, I, D, L, RP
	Ataques dirigidos				P, S, C, RL, I, D, L, RP	P, S, C, RL, I, D, L, RP			P, S, C, RL, I, D, L, RP	
. . . . .										
Escuchas / Interceptación	Man in the middle	P, I, C, RL, D, L, RP		P, I, C, RL, D, L, RP					P, I, D, L, RP	P, I, C, RL, D, L, RP
	Secuestro de protocolos de comunicación de IoT	P, I, C, RL, D, L, RP			P, I, C, RL, D, L, RP		P, I, C, RL, D, L, RP		P, I, D, L, RP	P, I, C, RL, D, L, RP
	Interceptación de la información	P, I, C, RL, D, L, RP		P, I, C, RL, D, L, RP					P, I, D, L, RP	P, I, C, RL, D, L, RP
. . . . .										
Cortes	Corte de red			C, RL, D, L, RP	C, RL, D, L, RP					
	Fallo de los dispositivos	C, RL, D, L, RP								C, RL, D, L, RP
	Perdida de soporte del dispositivo	C, RL, D, L, RP	C, RL, D, L, RP	C, RL, D, L, RP	C, RL, D, L, RP	C, RL, D, L, RP	C, RL, D, L, RP	C, RL, D, L, RP	C, RL, D, L, RP	C, RL, D, L, RP

Dimensiones: P:Privacidad; S:Salud; C:Confiabilidad; RL:Resiliencia; I:Integridad; D:Disponibilidad; RP:Responsabilidad; L:Legalidad



Figura 4. Ocurrencia y degradación por los tipos de amenazas en la herramienta eMARISMA.

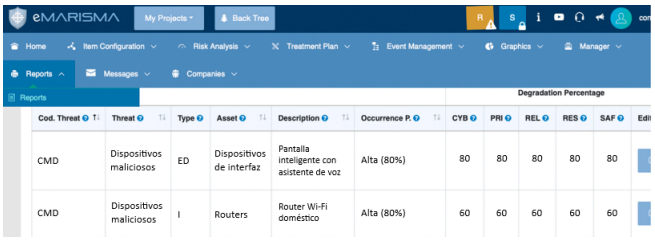


Figura 5. Procentaje de degradación x Dimensiones x Activos x Amenazas en la herramienta eMARISMA.

menos probable, y un porcentaje de degradación del 100 % debido a que se considera que la privacidad de los usuarios en infraestructuras IoT es de suma importancia y un ataque a este tendría un impacto muy elevado.

Inicialmente, la herramienta carga valores predeterminados basados en análisis previos y aprendizaje propio, pero estos pueden ser ajustados por expertos en seguridad. Para amenazas irrelevantes o que no son adecuadas para estos entornos, se pueden asignar valores bajos o nulos. Los valores establecidos afectan todas las relaciones entre amenazas y activos vinculados.

V-D. Análisis de riesgos para el caso de estudio

El siguiente paso en el análisis de riesgos es evaluar el impacto de las amenazas en los activos, considerando sus dimensiones. Esto permite calcular el daño total sufrido cuando una amenaza afecta el sistema, basado en el daño por dimensión. La herramienta visualiza la relación entre amenazas, activos y dimensiones de la Tabla II con los valores asignados previamente en la Figura 5.

Los valores predeterminados de degradación por dimensiones se establecen según el porcentaje definido en el paso anterior en la Figura 4, y por la relación entre activos, amenazas y dimensiones definidas en el patrón (ver Tabla II). Pueden asignarse automáticamente según el conocimiento de la herramienta o ser ajustados por expertos en seguridad, quienes pueden modificarlos según su criterio.

V-E. Resultados del análisis de riesgos

Para concluir el análisis de riesgos, se realiza una verificación interna mediante una lista de verificación de seguridad,

Tabla III
ACTIVOS DEL CASO DE ESTUDIO DEFINIDOS EN PATRON MARISMA-IOT.

Patrón MARISMA-IoT		**Patrón instanciado**
Familias de Activos	**Tipos de Activos**	**Activos del Caso de Estudio**
Dispositivos IoT	Hardware	Microcontrolador embebido de los dispositivos IoT
	Software	Firmware interno de los dispositivos IoT
	Sensores	Sensor de temperatura y humedad
	Actuadores	Motor de persianas automáticas
	S/A	Termostato inteligente
Ecosistema IoT	Dispositivos de interfaz	Pantalla inteligente con asistente de voz
	Dispositivos de gestión	Concentrador (hub) de hogar inteligente
	Sistemas embebidos	Mini-computador Raspberry Pi
Comunicaciones	Redes	Red Wi-Fi doméstica y Red Zigbee
	Protocolos	Protocolo MQTT
Infraestructura	Routers	Router Wi-Fi doméstico
	Gateways	Pasarela (gateway) multi-protocolo
	Fuentes de alimentación	Batería recargable de litio en dispositivos IoT compatibles
	Activos de seguridad	Cifrado WPA2 en la red Wi-Fi y credenciales seguras
Plataforma y Backend	Servicios Web	Servicio web de automatización IFTTT
	Infraestructura y servicios cloud	Plataforma en la nube AWS IoT
Toma de decisiones	Mineria de datos	Algoritmo de inteligencia artificial en la nube
	Procesamiento y computación de datos	Computación en el borde (edge)
Aplicaciones y servicios	Analisis y visualización de datos	Dashboard de Grafana para IoT
	Gestión de dispositivos y redes	Aplicación de control domótico unificado
	Uso de dispositivo	Automatización basada en contexto de uso
Información	En reposo	Datos históricos almacenados
	En transito	Mensaje IoT en transmisión
	En uso	Dato utilizado en tiempo real
Audio/Visual	Sistemas de streming digital	Televisor inteligente (Smart TV)

Tabla IV
SEGMENTO DE AMENAZAS DEL CASO DE ESTUDIO.

Familia Amenazas	Tipo Amenazas	Amenaza Caso Estudio
Actividad maliciosa y Abuso	Malware	Dispositivos infectados con programas maliciosos
	Kits de explotación	Exploits en los firmwares, servicios locales y cloud, etc.
	Ataques dirigidos	Ataque sofisticado sobre el hub domótico, dispositivo IoT, servicio local, etc.
	Dispositivos maliciosos	Dispositivos falsos de marca desconocida
	Ataques privacidad	Grabaciones de conversaciones privadas por parte de los dispositivos inteligentes, imagenes, etc.
	Modificación información	Modificación de los datos de los sensores, scripts de ejecución de acciones, etc.
Escuchas clandestinas, Interceptación y Secuestro	Man in the middle	Interceptación de comunicación con un smart lock, sincronizaciones entre dispositivos, etc.
	Secuestro de los protocolos de comunicación IoT	Falsificación de comandos entre dispositivos
	Interceptación de la información	Sniffer Wi-Fi, Bluetooth, Zigbee, etc.

basada en un patrón que define dominios, objetivos de control y controles. Su propósito es evaluar la cobertura de seguridad, identificando controles implementados, pendientes o descartados, y proporcionando una visión general de fortalezas y debilidades para la toma de decisiones.

La herramienta utilizada ofrece un panel con seguimiento en tiempo real de la cobertura de controles organizados por sus correspondientes dominios, visualizado gráficamente en la Figura 6 donde se puede observar, por ejemplo, que se tiene una cobertura del 50 % en los controles de perímetro de seguridad física. También permite analizar el nivel de seguridad mediante diagramas Kiviat por auditoría general, dominio y objetivos de control junto con mapas de calor con los riesgos inherentes y residuales al aplicar los controles, tal y cómo se puede ver en la Figura 7. Además, muestra información detallada sobre el riesgo, incluyendo niveles por tipo de activo y nivel de amenaza como es visible en la Figura 8, calculando el valor promedio de riesgo para cada tipo de activo, como por ejemplo, el activo *Software* tiene asignado un Riesgo Medio del 25 %, un riesgo menor del 5 % y un riesgo mayor del 70 %. De esta manera podemos visualizar el estado general de la seguridad de toda la infraestructura IoT, desde los riesgos residuales, hasta los niveles de cobertura obtenidos con los diferentes controles aplicados, dando así una visión transversal que permite una toma de decisiones más precisa a la hora de la gestión de riesgos.

VI. CONCLUSIÓN Y TRABAJO FUTURO

En este trabajo se ha definido un patrón para entornos IoT basado en MARISMA, con cuatro catálogos taxonómicos (controles, activos, amenazas y dimensiones) basados en estándares internacionales. Incluye matrices de dependencia que facilitan la reutilización y adaptación del conocimiento. Su integración en MARISMA fue validada en casos prácticos.

A medio plazo, se plantea que a partir del patrón desarrollado generar versiones más específicas adaptadas a distintos sectores o incluso ajustarlo a las necesidades particulares de estos, ya que gracias a las capacidades de herencia y reutilización que ofrece MARISMA, se pueden crear subpatrones orientados a contextos IoT concretos, como fábricas, oficinas o ciudades inteligentes, permitiendo un análisis de riesgos más preciso con un menor consumo de recursos.

VII. AGRADECIMIENTOS

Este trabajo se ha desarrollado gracias a la financiación proporcionada por los siguientes proyectos:

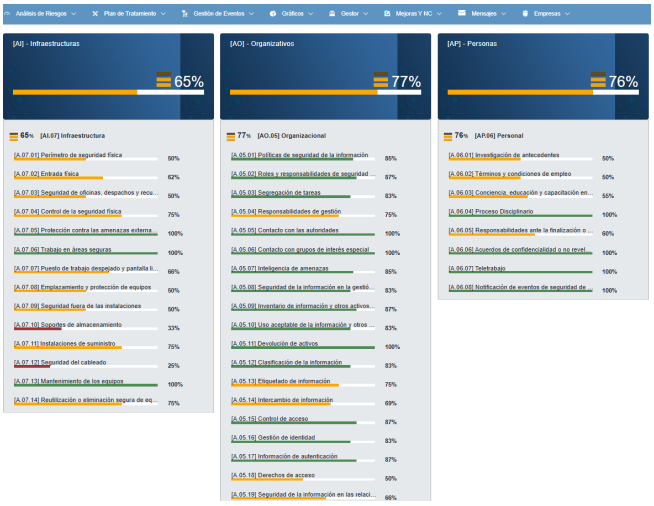


Figura 6. Panel de control con los niveles de cobertura de los controles del caso de estudio.

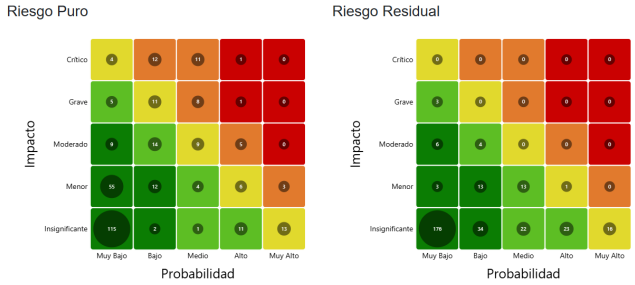


Figura 7. Mapa de calor con riesgos inherentes y residuales.

Proyecto Di4SPDS (CHIST-ERA grant- PCI2023145980-2) financiado por MCIN/AEI y cofinanciado por la Unión Europea, KOSMOS-UCLM (PID2024-155363OB-C44) financiado por MCIN/AEI/10.13039/501100011033; ALBA-UCLM (TED2021-130355B-C31) financiado por CIN/AEI/10.13039/501100011033/ Unión Europea NextGenerationEU/PRTR, y MESIAS (2022-GRIN-34202) financiado por FEDER.

Referencias

[1] M. V. Moreno, B. Úbeda, A. F. Skarmeta, y M. A. Zamora, "How can we tackle energy efficiency in IoT-based smart buildings?,"*Sensors*, vol. 14, no. 6, pp. 9582-9614, 2014. doi: https://doi.org/10.3390/s140609582.

[2] M. García-Monge, B. Zalba, R. Casas, E. Cano, S. Guillén-Lambea, B. López-Mesa, e I. Martínez, "Is IoT monitoring key to improve building energy efficiency? Case study of a smart campus in Spain,"*Energy and Buildings*, vol. 285, p. 112882, 2023. doi: https://doi.org/10.1016/j.enbuild.2023.112882.

[3] L. Cambosuela, M. Kaur, y R. Astya, "The vulnerabilities and risks of implementing Internet of Things (IoT) in cyber security,ïn *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2024, pp. 1-5. doi: https://doi.org/10.1109/ICRITO61523.2024.10522460.

[4] M. A. Razzaq, S. H. Gill, M. A. Qureshi, y S. Ullah, "Security issues in the Internet of Things (IoT): A comprehensive study,"*International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 6, 2017. doi: http://dx.doi.org/10.14569/IJACSA.2017.080650

[5] T. E. Abioye, O. T. Arogundade, S. Misra, K. Adesemowo, y R. Damaševičius, "Cloud-based business process security risk management: A systematic review, taxonomy, and future directions,"*Computers*, vol. 10, no. 12, p. 160, 2021. doi: https://doi.org/10.3390/computers10120160.

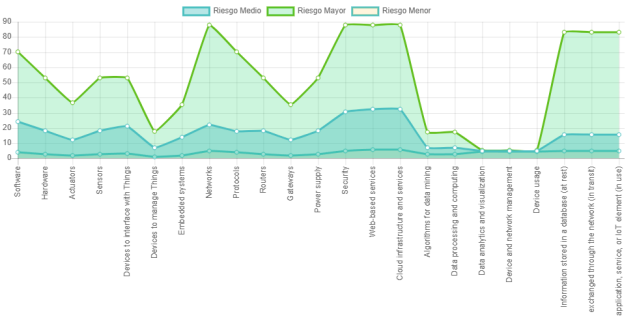


Figura 8. Riesgo calculado por tipo de activo utilizando la herramienta eMARISMA para el caso de estudio.

[6] M. N. Aleksandrov, V. A. Vasiliev, y S. V. Aleksandrova, "Implementation of the risk-based approach methodology in information security management systems,ïn *2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, Yaroslavl, Russian Federation, 2021, pp. 137-139. doi: https://doi.org/10.1109/ITQMIS53292.2021.9642767.

[7] L. E. Sánchez, A. Santos-Olmo, D. G. Rosado, C. Blanco, M. A. Serrano, H. Mouratidis, y E. Fernández-Medina, "MARISMA: A modern and context-aware framework for assessing and managing information cybersecurity risks", *Computer Standards & Interfaces*, vol. 92, p. 103935, 2025. doi: https://doi.org/10.1016/j.csi.2024.103935.

[8] D. G. Rosado, J. Moreno, L. E. Sánchez, A. Santos-Olmo, M. A. Serrano, y E. Fernández-Medina, "MARISMA-BiDa pattern: Integrated risk analysis for big data", *Computers & Security*, vol. 102, p. 102155, 2021. doi: https://doi.org/10.1016/j.cose.2020.102155.

[9] D. G. Rosado, A. Santos-Olmo, L. E. Sánchez, M. A. Serrano, C. Blanco, H. Mouratidis, y E. Fernández-Medina, "Managing cybersecurity risks of cyber-physical systems: The MARISMA-CPS pattern", *Computers in Industry*, vol. 142, p. 103715, 2022. doi: https://doi.org/10.1016/j.compind.2022.103715.

[10] D. G. Rosado, L. E. Sánchez, Á. J. Varela-Vaca, A. Santos-Olmo, M. T. Gómez-López, R. M. Gasca, y E. Fernández-Medina, "Enabling security risk assessment and management for business process models", *Journal of Information Security and Applications*, vol. 84, p. 103829, 2024. doi: https://doi.org/10.1016/j.jisa.2024.103829.

[11] ISO/IEC 27001:2023, "Information security, cybersecurity and privacy protection — Information security management systems — Requirements", International Organization for Standardization, 2023. doi: https://www.iso.org/standard/27001

[12] ISO/IEC 27005:2022, "Information security, cybersecurity and privacy protection — Guidance on managing information security risks", International Organization for Standardization, 2022. doi: https://www.iso.org/standard/80585.html

[13] European Union Agency for Network and Information Security, "Baseline Security Recommendations for IoT in the context of Critical Information Infrastructures", ENISA, 2017. doi: https://www.enisa.europa.eu/publications/baseline-security-recommendations-for-iot

[14] European Union Agency for Cybersecurity, "ENISA Threat Landscape for Smart Home Environments", ENISA, 2020. doi: https://www.enisa.europa.eu/publications/enisa-threat-landscape-for-smart-home-environments

[15] K. A. Megas, M. Fagan, K. Scarfone y M. Smith, "IoT Device Cybersecurity Capability Core Baseline", NIST Interagency or Internal Report (NISTIR) 8259A, National Institute of Standards and Technology, 2020. doi: https://doi.org/10.6028/NIST.IR.8259A

[16] K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams y A. Hahn, "Guide to Industrial Control Systems (ICS) Security", NIST Special Publication 800-82 Revision 2, National Institute of Standards and Technology, 2015. doi: https://doi.org/10.6028/NIST.SP.800-82r2

[17] J. Boyens, C. Paulsen, R. Moorthy y N. Bartol, "Supply Chain Risk Management Practices for Federal Information Systems and Organizations", NIST Special Publication 800-161, National Institute of Standards and Technology, 2015. doi: https://doi.org/10.6028/NIST.SP.800-161

Análisis comparativo de las capacidades de SIDS en un escenario SmartHome real

J. Muñoz-Calle*, J. Diaz-Verdejo†, R. Estepa Alonso*, A. Estepa Alonso*
* Dpt. Ingeniería Telemática, Escuela Superior de Ingenieros, Univ. de Sevilla
C/ Camino de los Descubrimientos s/n, 41092 Sevilla (Spain)
E-mail: {fjmc,rafaestepa,aestepa}@us.es
† Dpt. Teoría de Señal, Telemática y Comunicaciones, CITIC, Univ. de Granada
C/ Periodista Daniel Saucedo Aranda, s/n, 18071 Granada (Spain)
E-mail: jedv@ugr.es

***Resumen*—Cada vez son más los dispositivos desplegados en entornos SmartHome residenciales para conseguir funcionalidades de control relativamente simples. Su bajo coste y facilidad de uso propician una fuerte expansión, lo que a su vez representa un desafío desde el punto de vista de la ciberseguridad, aumentando significativamente la exposición de las redes residenciales. La utilización de sistemas de detección de intrusiones adaptados al contexto podría mejorar la seguridad. Este trabajo estudia la idoneidad de los IDS para la detección de ciberataques en un escenario tipo SmartHome real, utilizando tanto detectores de dominio público como comerciales. Para ello se analizan trazas reales disponibles con 3 IDS ampliamente utilizados: Snort, Palo Alto NGFW y FortiGate. Los resultados obtenidos ilustran la imposibilidad de su despliegue en su configuración por defecto, planteando algunas cuestiones relativas al rendimiento y la dificultad de comparar sus rendimientos debido al punto de operación elegido en los equipos comerciales frente a Snort.**

***Index Terms*—seguridad en SmartHome, detección de ciberataques, SIDS**

Tipo de contribución: *Investigación original*

I. Introducción

Los dispositivos inteligentes han ido ganando presencia en los entornos residenciales a un ritmo acelerado en los últimos años. Cada vez más hogares incorporan equipos IoT (Internet of Things) de reducidas prestaciones para el control de iluminación, climatización, seguridad o entretenimiento, entre otros. Sin embargo, esta diversidad de dispositivos proviene de múltiples fabricantes y opera bajo distintos protocolos, tanto estándares como propietarios. Esta falta de uniformidad dificulta la integración eficiente de dispositivos, comprometiendo tanto la experiencia del usuario como la seguridad de las comunicaciones en estos entornos.

Actualmente, el mercado de SmartHome tiende a converger en torno a plataformas como Amazon Alexa, Google Home y Tuya, que centralizan la gestión de dispositivos a través de servicios en la nube. No obstante, el uso de protocolos cifrados y propietarios introduce nuevas barreras para la interoperabilidad y el control de los dispositivos. De hecho, esta es una de las grandes limitaciones existentes en la actualidad, ya que existen múltiples estándares, algunos propietarios, que resultan incompatibles entre sí o que presentan serios problemas de seguridad en la integración y operación. Además, la dependencia de las plataformas centralizadas plantea serios desafíos en términos de privacidad y ciberseguridad, ya que la comunicación con servidores externos aumenta la exposición a ataques y la posibilidad de fugas de datos sensibles. En este contexto, resulta fundamental analizar el impacto de estas limitaciones en la integración, operación y seguridad de los entornos SmartHome.

El escenario típico utiliza una red residencial (HAN, *Home Area Network*), en la que el router de acceso debe implementar NAT y los dispositivos *SmartHome* acceden a servicios en la nube a través de dicho router utilizando diversos mecanismos para atravesar el NAT que posibiliten su control remoto. Por otra parte, los dispositivos están diseñados para usar una pila de protocolos reducida con limitaciones de cómputo y memoria. Adicionalmente, el usuario residencial típico carece de la experiencia y los conocimientos necesarios para securizar adecuadamente la red y la mayoría de los dispositivos se proporcionan bajo el esquema de una instalación automatizada y simple para el usuario. El resultado es una fuerte exposición a ataques tanto desde equipos locales como remotos.

Así, en trabajos previos de los autores [1] se ha constatado la exposición, los numerosos protocolos propietarios y el aumento significativo del tráfico, tanto interno como hacia el exterior, sin justificación aparente. Otros autores (p.e. [2] [3]) evidencian el incremento en el nivel de riesgo y la falta de medidas de seguridad adaptadas al contexto.

Una de las herramientas de seguridad utilizadas habitualmente son los sistemas de detección de intrusiones (IDS, por sus siglas en inglés *Intrusion Detection Systems*), por lo que se podría plantear su utilización en este escenario. La aproximación más utilizada para el despliegue de IDS son los IDS basados en firmas (SIDS). En trabajos previos [4] hemos evidenciado la necesidad de adecuar las firmas utilizadas por los SIDS al contexto y escenario concreto y el fuerte impacto de estas en las tasas de detección y de falsos positivos. Sin embargo, el tráfico de los dispositivos *SmartHome* presenta algunas características diferenciadoras [5] que podrían influir en la capacidad de detección de los IDS estándar. Así, además de la ya mencionada pila de protocolos y capacidades reducidas, se pueden identificar 3 patrones típicos de tráfico en los dispositivos *SmartHome* asociados, respectivamente, a actualizaciones periódicas, a la respuesta a eventos o a intercambios de datos. Estos patrones difieren de los habituales en redes residenciales que no incorporen dispositivos *SmartHome*.

Por tanto, cabe plantearse la utilidad del despliegue de SIDS en estos escenarios y sus capacidades efectivas, así como la idoneidad de las firmas existentes. Para ello, nos proponemos

evaluar el rendimiento de algunas herramientas SIDS tanto gratuitas como de pago. En particular, consideramos un SIDS gratuito de amplio uso, como es *Snort* [6] con firmas *Talos* y *ETOpen*, y su comparación con dos sistemas comerciales: *FortiGate* [7] y *Palo Alto NGFW* [8]. Ambos dispositivos corresponden a cortafuegos de nueva generación, con múltiples funcionalidades de seguridad incorporadas. De todas ellas, en el presente trabajo usaremos su capacidad de identificación por firmas. Por este motivo, en el resto del texto nos referimos a ambos dispositivos como SIDS. La evaluación se realiza sobre un escenario real en uso.

En la literatura hemos encontrado trabajos (véase Apartado II) que evalúan el rendimiento de sistemas IDS en escenarios *SmartHome* en términos de coste computacional, así como otros que proponen y evalúan IDS basados en anomalías. Sin embargo, no hemos podido encontrar ninguna contribución que aborde realmente la adecuación de los SIDS y firmas existentes.

El resto del artículo se estructura como sigue. En el Apartado II se describen brevemente trabajos previos relacionados con la detección de intrusiones en IoT y *SmartHome*, motivándose las contribuciones de este trabajo. En el Apartado III se describen el escenario real utilizado y los datos adquiridos en el mismo para la experimentación. También se describen las herramientas SIDS a evaluar. A continuación, en el Apartado IV se presentan los resultados de detección obtenidos con las diferentes configuraciones. Estos resultados son comparados y analizados en el Apartado V. Finalmente, en el Apartado **??** se presentan las conclusiones y el trabajo futuro.

II. Antecedentes

En un reciente trabajo de revisión [3] se identifican los 6 desafíos más relevantes que plantea la securización de entornos *SmartHome*, presentando las líneas de avance más relevantes en la actualidad. Entre ellas se especifica la identificación de soluciones para la detección de intrusiones en *SmartHome* como una de las líneas de trabajo más activas desde el punto de vista de la investigación.

Los trabajos descritos en la literatura en relación al uso de IDS en entornos SmartHome suelen poner el foco en cuestiones relacionadas, por una parte, con el coste computacional y la forma de despliegue [5], y, por otra parte, con el desarrollo y adaptación de detectores basados en anomalías [9]. En cualquier caso, sorprendentemente, pocos trabajos usan datasets con tráfico IoT y menos aún con tráfico *SmartHome* (p.e. [10]). Es más, se suelen usar datasets simulados [9], evaluando la capacidad con tipos de ataques concretos (p.e. DoS, buffer overflow, etc.). Por tanto, finalmente operan sobre datasets y escenarios que no son realmente *SmartHome*. Este problema está bastante extendido, limitando la validez de las propuestas y análisis realizados.

Así, a modo de ejemplo, [5] plantea la detección de tráfico anómalo en *SmartHome*, pero se centra en dónde ubicar el detector. Para la evaluación de su propuesta, utiliza un escenario simulado. Por su parte, [10] se centra en comparar el rendimiento de varios detectores desde el punto de vista de su coste computacional.

El ámbito más activo es el de la detección basada en anomalías, con soluciones mayoritariamente basadas en machine learning, especialmente aplicando técnicas de *deep learning* [11]. En estos casos, típicamente se aborda el problema desde una aproximación 'ciega': se entrenan y comparan los sistemas sin ningún tipo de interpretación/adaptación de los datasets/parámetros al escenario concreto [12], seleccionándose el que funciona mejor según la métrica elegida. Otros trabajos plantean aproximaciones basadas en otro tipo de técnicas de aprendizaje automático, incluyendo combinaciones de técnicas, como “isolation forest” y *clustering* [9].

En relación a nuestro análisis, el único trabajo que hemos identificado que plantea la comparación de SIDS en entornos SmartHome es [10]. Sin embargo, aunque compara el rendimiento de *Snort, Suricata* y *Bro*, lo hace en términos de CPU/RAM, no considerando ningún indicador relacionado con las tasas de FP (False Positive) o TP (True Positive). Por otra parte, no se analiza tráfico SmartHome ni se considera un escenario de este tipo. La evaluación se realiza a partir de capturas de datos obtenidas de malware-traffic-analysis.net.

III. Escenario experimental

A continuación se describe brevemente el escenario real utilizado para la experimentación y el dataset capturado en el mismo. También se presentan brevemente los sistemas SIDS que han sido aplicados para analizar dicho dataset.

III-A. Red SmartHome y dataset

La red supervisada, denominada *Smart Home TB*[1], corresponde a la desplegada a demanda en una vivienda unifamiliar con varias plantas que está siendo utilizada como plataforma experimental por nuestro equipo de investigación. La arquitectura de la red responde al modelo propuesto en [13] y se encuentra segmentada en varias subredes interconectadas mediante *switches*. También se incorporan varios puntos de acceso WIFI operando con el mismo SSID. Una descripción más detallada se presenta en [1]. Es relevante indicar que las únicas modificaciones sobre la red previamente existente han consistido en la sustitución de algunos equipos de red por versiones gestionables para posibilitar la adquisición de los datos.

En la Fig. 1 se muestra un esquema con los equipos y tecnologías presentes de forma permanente en el despliegue. A estos hay que sumar los equipos Wifi de los usuarios (teléfonos móviles, portátiles, etc.), que se conectan a la red a demanda. Los equipos de tipo SmartHome (recuadros en azul) se encuentran agrupados en función de su modo de conexión (cableado/inalámbrico), indicándose la tecnología SmartHome nativa de cada dispositivo. Así, se incorporan las siguientes: eWelink (1 disp.), merros (3 disp.), tuya (1 disp.), tplink (1 disp.), espHome (2 disp.) y alexa (1 disp.). Adicionalmente, se incluyen dispositivos que utilizan protocolos estándar (las cámaras), así como otros que operan sobre protocolos propietarios no documentados. En total se consideran 20 dispositivos SmartHome. Adicionalmente, el escenario incluye un *Home server* sobre el que se ejecuta el software de domótica *Home Assistant*[2] (HA). El objetivo de este servidor es centralizar la gestión de todos los dispositivos SmartHome localmente, aunque no ha sido posible incorporar algunos de ellos por no estar disponibles los módulos adecuados (Fig. 2). En

[1] https://dtstc.ugr.es/neus-cslab/recursos/monitorizacion-smarthome/
[2] https://www.home-assistant.io/

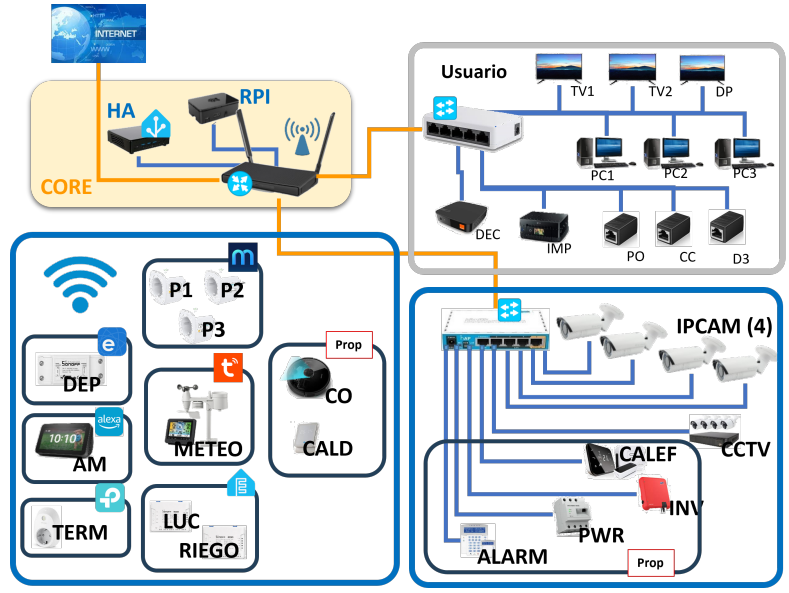


Figura 1. Equipos y tecnologías en la red real utilizada para la adquisición de datos [1].

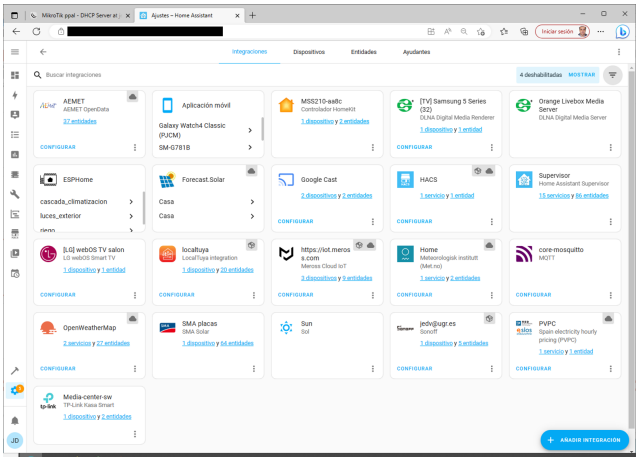


Figura 2. Plug-ins incluidos en el *Home server* para el control local de los dispositivos.

cualquier caso, además del control local en los casos en los que era posible, se ha mantenido para la experimentación la operación nativa de los dispositivos, típicamente basada en la nube/servidores de cada tecnología.

La configuración de los equipos más críticos y complejos (sistemas de seguridad, caldera, placas fotovoltaicas, etc.) ha sido realizada por instaladores autorizados. Para el resto de equipos se han seguido rigurosamente las instrucciones proporcionadas por el fabricante. Asimismo, el router ha sido configurado siguiendo las indicaciones de mejores prácticas, especialmente en lo que respecta al cortafuegos y servicio NAT.

III-A1. Dataset: El dataset utilizado en el presente trabajo corresponde a una de las campañas de adquisición de datos[3] realizadas en el escenario previamente descrito. En particular, es el adquirido durante la campaña 1, en la que se ha capturado el tráfico de todos los servicios y dispositivos asociados a SmartHome durante 30 días. Este tráfico se ha adquirido a partir de la operación real del escenario, capturando las cargas útiles completas y seleccionando los paquetes con origen o destino en alguno de los dispositivos SmartHome. A fin de evitar un volumen de datos excesivo, se ha filtrado el tráfico RSTP de los *stream* de las cámaras.

El resultado es un dataset multiservicio que incluye tráfico de diversos protocolos y equipos SmartHome, incorporándose tanto servicios propietarios de fabricantes como servicios estandarizados o de amplio uso. Desafortunadamente, este carácter propietario de algunos protocolos ha imposibilitado hasta la fecha la anonimización completa del dataset, por lo que no se encuentra aún públicamente disponible. Por otra parte, es importante señalar que el dataset no se encuentra etiquetado.

Las características más relevantes del dataset se muestran en la Tabla I. Una parte de este dataset fue utilizado en nuestro trabajo previo [1], analizándose únicamente los 3 primeros días de la captura. En el presente trabajo se considera todo el dataset, centrándonos en la detección de ataques mediante SIDS. En cualquier caso, se extrae la jerarquía de protocolos mediante *Wireshark*, obteniéndose los resultados mostrados en la Fig. 3. Nótese el elevado porcentaje de tráfico TCP no identificado así como el significativo volumen de datos asociados al protocolo *TP-Link SmartHome Protocol*, pese a la presencia de un único dispositivo de este tipo. Asimismo, el tráfico ARP representa un volumen a priori excesivo en el escenario considerado. En cualquier caso, el análisis de los flujos/sesiones como tal queda fuera del objeto de este trabajo.

III-B. Herramientas de detección

Para la identificación de los ataques, las trazas recogidas (formato pcap) han sido analizadas mediante tres detectores extensamente utilizados, uno de ellos *Snort* [6], de dominio público, y otros dos de ámbito comercial, *FortiGate* [7] y *Palo Alto NGFW* [8].

[3]https://dtstc.ugr.es/neus-cslab/recursos/ds-iot/

Duración	30 días
Vol (GB)	4.7
N. IP Locales	62
N. IP total	611
N. Paq.	17786742
N. Endpoints (TCP/UDP)	32023 / 89891

Tabla I
CARACTERÍSTICAS MÁS RELEVANTES DEL DATASET UTILIZADO.

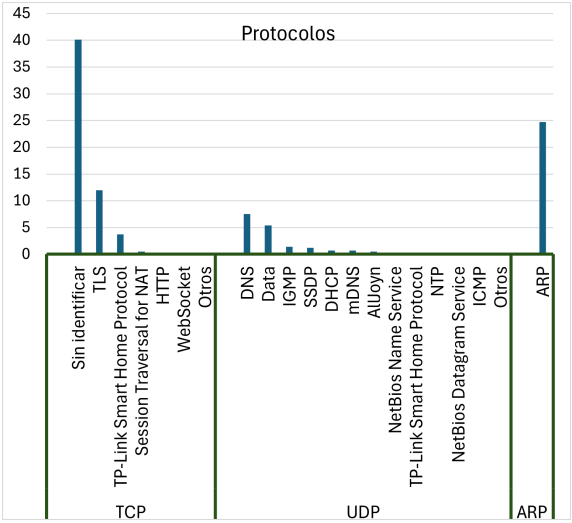


Figura 3. Jerarquía de protocolos en el dataset (% paquetes).

Snort corresponde a un software SIDS genérico de amplia implantación. Para dar soporte a los distintos paquetes de firmas empleados, han sido usadas las versiones de *Snort* 2.9 y 3.0. La herramienta ha sido instalada en un SO *Linux Debian 12.9 (bookworm)*, manteniendo la configuración por defecto de los preprocesadores. El tráfico ha sido analizado de forma pasiva, mediante la lectura directa del fichero de trazas por parte de *Snort*.

Por su parte, *FortiGate* y *Palo Alto NGFW* corresponden a cortafuegos de nueva generación desarrollados por las empresas Fortinet y Palo Alto Networks, respectivamente, especializadas en el desarrollo de dispositivos y servicios de ciberseguridad. Ambas son empresas líderes a nivel global en soluciones de ciberseguridad completas, ofreciendo un amplio ecosistema de productos, siendo *FortiGate* y *Palo Alto NGFW* sus líneas de cortafuegos para la protección avanzada contra ataques en red (IDS) y filtrado de tráfico en tiempo real.

De los distintos modelos con que son distribuidos estos cortafuegos, para desarrollar nuestro estudio ha sido usado el modelo basado en tecnología de virtualización KVM (Kernel-based Virtual Machine), desplegado sobre un sistema operativo anfitrión Linux Debian 12.9 soportado en un servidor HP Proliant ML110 Gen11.

En el caso de *FortiGate*, se ha desplegado el modelo `FortiGate VM64-KVM`, con el sistema operativo `FortiOS 7.4.4`. Para satisfacer los requisitos mínimos que garantizasen un funcionamiento adecuado, se dotó a la máquina virtual del cortafuegos con 8 GB de memoria RAM y 2 CPUs virtuales. Por su parte, para el cortafuegos *Palo Alto NGFW* se ha usado el modelo `PA-VM VM-SERIES-8 KVM`, con el sistema operativo `PAN-OS 11.1.4`. Para este dispositivo, con una demanda de recursos más exigente, se dotó a la máquina virtual con 16 GB de memoria RAM y 8 CPUs virtuales. En ambos dispositivos, el proceso de instalación, configuración y valoración de los resultados se ha realizado con el apoyo y supervisión de los equipos de ingeniería de soporte del fabricante.

Dado que el objeto principal de estos dispositivos comerciales es actuar como cortafuegos perimetrales, analizando el tráfico en tiempo real, a diferencia de *Snort* no están preparados para acceder y analizar de manera directa un fichero de trazas (pcap). No obstante, los modelos virtuales empleados disponen de la capacidad de usar interfaces en un modo especial de funcionamiento, denominado modo *sniffer*, bajo el cual pueden recibir tráfico bidireccional generado desde el sistema operativo anfitrión. Aprovechando esta posibilidad, el tráfico capturado fue enviado a la interfaz en modo *sniffer* de cada cortafuegos mediante su reproducción, desde el sistema operativo anfitrión, con la herramienta `tcpreplay-edit` integrada en el paquete opensource de utilidades para tráfico de red `tcpreplay` [14].

Resulta relevante advertir que, difiriendo también con *Snort*, estos dispositivos, habitualmente ubicados en nodos por los que atraviesa todo el tráfico de la red, están capacitados para analizar mensajes a nivel de red, pero no a nivel de enlace. Consecuentemente, si en el fichero de tráfico a analizar existiesen ataques hacia dispositivos de este nivel, como conmutadores o puntos de acceso, en general no serán detectados.

Aunque ambos dispositivos disponen de una potente interfaz de usuario, para extraer los datos de manera procedimental se usaron sus ficheros de registro (`logs`), identificando las alertas de ataque mediante el análisis de los campos de cada entrada.

III-C. Firmas de detección

Para detectar los ataques y generar las alertas, cada detector usa un juego de firmas o firmas específicas. En la Tabla II se resumen las características de los paquetes de firmas que han sido empleados.

En el caso de Snort, se han usado dos paquetes de firmas diferentes:

a) Talos, desarrolladas por el grupo de inteligencia de amenazas Cisco Talos o, formalmente, VRT (Vulnerability Research Team) [15]. Se han empleado las dos modalidades gratuitas en que se distribuye el paquete, Community y Registered. Las firmas Community, usadas por defecto por Snort, son el conjunto de firmas desarrolladas y mantenidas por la comunidad de usuarios, con una frecuencia de actualización variable. Las firmas Registered, ofrecidas bajo registro gratuito en el portal de Snort, son firmas certificadas por el grupo Talos, actualizadas cada 30 días, conteniendo todas las firmas comunitarias, ademas de un amplio conjunto de firmas adicionales con mayor refinamiento para mejorar las detecciones y reducir la tasa falsos positivos.
b) ETOpen, generadas por el grupo de investigación en ciberseguridad Proofpoint ET (Emerging Threats) [16]. Distribuidas de manera gratuita y en código abierto, son mantenidas por la comunidad, actualizándose de forma casi diaria. El paquete ha sido usado en dos modalidades. La primera, que hemos denominado simplemente

Detector	Paquete	#Firmas	Fecha
Snort	Talos Community	4022	30/05/2024
	Talos Registered	45557	
	ETOpen	35763	29/05/2024
	ETOpen Optimizada	47830	
FortiGate		12303	04/07/2024
Palo Alto NGFW		21075	

Tabla II
DETECTORES Y FIRMAS

RuleSet	Talos Community	Talos Registered	ETOpen	ETOpen optimizada
RS1	✓			
RS2	✓		✓	
RS3		✓	✓	
RS4		✓		✓

Tabla III
RULESETS POR COMBINACIÓN DE LOS PAQUETES DE FIRMAS DE SNORT

ETOpen, corresponde a la versión original en que es distribuido. La segunda, que hemos identificado como ETOpen optimizada, ha sido el resultado de modificar la versión anterior, activando (descomentando) todas las firmas del paquete, salvo las marcadas como obsoletas (ET DELETED), con el propósito de maximizar la capacidad de detección a pesar del posible incremento en falsos positivos.

Con el objetivo de analizar la capacidad de detección conjunta de estos paquetes de firmas de Snort, estos han sido agrupados en cuatro conjuntos de firmas o RS (RuleSets), de forma incremental. Las combinaciones de firmas se resumen en la Tabla III .

Para los cortafuegos FortiGate y Palo Alto NGFW, se extrajeron las detecciones realizadas por su motor IPS (Intrusion Prevention System), dotado de los distintos paquetes de firmas disponibles para estos productos, incluyendo las firmas para detección de amenazas OT (Operative Technology), IoT (Internet of Things) e ICS (Industrial Control Systems), actualizados a la fecha de realización del análisis (23 de abril de 2024), con un total de 12 303 y 21 075 firmas, respectivamente.

IV. METODOLOGÍA Y RESULTADOS OBTENIDOS POR CADA SIDS

IV-A. Metodología

El tráfico capturado durante los 30 días fue analizado por las tres herramientas de detección, empleando sus respectivos conjuntos de firmas. Para evaluar el rendimiento de un detector, además del número de alertas, será necesario determinar qué porcentaje de ellas corresponden a alertas reales o TP. Sin embargo, esta cuestión resulta compleja, dado que el dataset utilizado se corresponde con un entorno real cuyo elevado volumen de datos (17 millones de paquetes) imposibilita la inspección manual completa. Por ello se ha realizado una aproximación para la clasificar las alertas como verdadero (TP) o falsos positivo (FP) en función de la clasificación otorgada previamente a la firma que generó la alerta.

Utilizaremos los siguiente criterios para realizar la clasificación de cada firma como TP:

- Inspección manual de los primeros tres días: Se ha utilizado el análisis manual de los flujos de tráfico asociados a cada SID durante los tres primeros días de captura. Este análisis permitió identificar diversos ataques, según lo descrito en estudios previos [1]. Por tanto, cualquier SID presente en la lista de firmas clasificadas como TP en estos días será considerado TP en el resto del análisis.
- Aproximación basada en el número de alertas: Para los SID no incluidos en la clasificación anterior, se ha empleado un enfoque basado en el volumen de alertas generadas por cada firma. La experiencia en el análisis de alertas indica que las firmas que producen un número excesivo de alertas tienden a ser falsos positivos (FP). Esto se debe a que, en entornos reales, la proporción de ataques es muy baja en comparación con el tráfico legítimo, lo que hace que las alertas TP sean significativamente menos numerosas. Aunque esta aproximación no es universalmente válida, ha sido verificada en nuestro estudio. La figura 4 muestra la distribución de alertas por SID en Snort, evidenciando que algunas firmas generan pocas alertas, mientras que otras producen volúmenes extremadamente altos (en escala logarítmica). Se han analizado individualmente las 62 firmas detectadas por Snort y los cortafuegos comerciales, contrastando el motivo de cada alerta con el tráfico y el contexto del experimento. Este proceso permitió confirmar que todas las firmas con menos de 100 alertas eran TP, mientras que aquellas con un mayor número de alertas resultaron ser FP.

En resumen, cada SID observado se ha clasificado como TP si se encuentra entre las alertas verificadas manualmente en los 3 primeros días o si generó menos de 100 alertas. Aunque esta clasificación no resulta exhaustiva, consideramos que los resultados que proporciona son representativos de la realidad, permitiendo observar tendencias confiables.

IV-B. Resultados: alarmas generadas y firmas activadas

Los resultados obtenidos se condensan en: Tabla IV, con las alarmas generadas por cada detector, así como los identificadores de las firmas que las generan (SID) para los dos períodos de tiempo señalados (30 días y los 3 primeros días), y Tabla V, con la misma información pero tan solo reflejando las alertas y SID consideradas TP según el criterio utilizado.

Durante los 30 días de análisis, Snort generó un volumen excepcionalmente alto de alertas, con un promedio diario de 3189 usando las firmas Talos Community (RS1) y 21 589 con las firmas RS4 (Talos Registered + ETOpenMax). En contraste, los cortafuegos comerciales produjeron significativamente menos alertas, con Palo Alto NGFW reduciéndolas en uno o dos órdenes de magnitud respecto a Snort. Palo Alto, sigue una aproximación similar, priorizando aún más la precisión sobre la cantidad de alertas. En cualquier caso el número de alertas diarias generadas resulta muy elevado para los tres SIDS evaluados.

Atendiendo a las firmas utilizadas por cada equipo para generar las alertas, podemos observar que Snort activa entre 10 (RS1) y 62 (RS4) firmas SID) diferentes. Palo Alto, con un total de 21 075 firmas, tan sólo activa 6 mientras que FortiGate, con 12 303 firmas activas sólo activa 3. Estos valores nos hacen pensar que las firmas de los equipos comerciales están especialmente diseñadas para evitar falsos positivos, contrariamente a lo que ocurre con las de Snort.

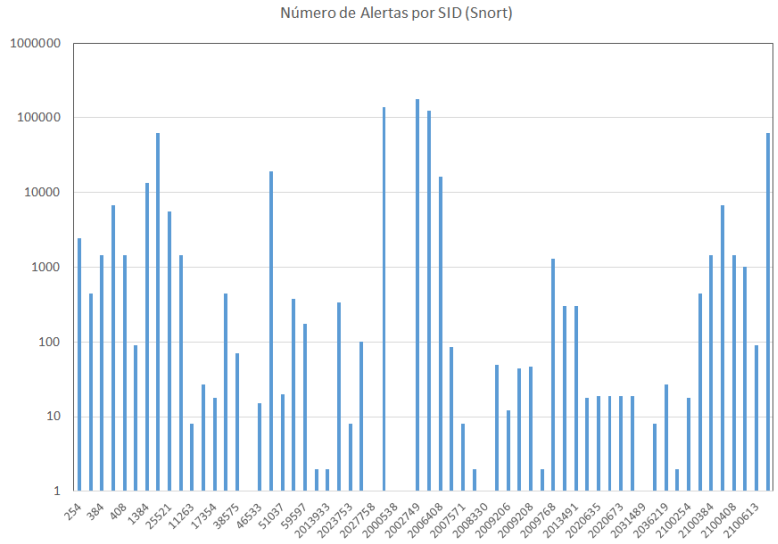


Figura 4. Número de Alertas por cada SID (Snort).

		3 días			30 días		
	Detector	**#SIDs**	**#Alertas**	**#Alertas/Día**	**#SIDs**	**#Alertas**	**#Alertas/Día**
Snort	Talos Community (RS1)	9	21666	7222	10	95668	3189
	Talos Registered	17	26042	8681	21	116076	3869
	ETOpen	5	32459	10820	7	138656	4622
	ETOpen Optimizada	32	71649	23883	41	531600	17720
	RS2	14	54125	18042	17	234324	7811
	RS3	22	58501	19500	28	254732	8491
	RS4	49	97691	32564	62	647676	21589
FortiGate		2	3655	1218	3	41667	1389
Palo Alto NGFW		3	3834	1278	6	26952	898

Tabla IV
ANÁLISIS DE DETECCIONES.

		3 días			30 días		
	Detector	**#SIDs**	**#Alertas**	**#Alertas/Día**	**#SIDs**	**#Alertas**	**#Alertas/Día**
Snort	Talos Community (RS1)	1	10	3	1	90	3
	Talos Registered	5	30	10	7	159	5
	ETOpen	2	7	2	4	13	0
	ETOpen Optimizada	18	56	19	22	493	16
	RS2	3	17	6	5	103	3
	RS3	7	37	12	11	172	6
	RS4	23	86	29	29	652	22
FortiGate		0	0	0	0	0	0
Palo Alto NGFW		1	3	1	4	14	0

Tabla V
ANÁLISIS DE VERDADEROS POSITIVOS (TP).

En cuanto a la capacidad de detección (TP), Snort ha mostrado una alta sensibilidad, activando múltiples firmas asociadas a ataques verificados manualmente en los tres primeros días. Por ejemplo, el primer día de la captura, a las 13:30, se dispara la alerta 2009582, desde recyber.net a PWR, lo cual indica que se realiza un escaneo.Posteriormente, se activaron las alertas 2000560 y 2008284, señalando intentos de establecer un túnel mediante conexiones HTTP en puertos cerrados. Más tarde, a las 5:54, las alertas 2034857 y 2023753 detectaron intentos de sesiones de escritorio remoto desde IP en listas negras, geolocalizadas en Bulgaria y Rusia. Estos intentos se repitieron desde la IP rusa a las 10:20 y 16:08. Las firmas Talos Registered, a pesar de tener un número de reglas similar a ETOpen Max, generaron significativamente menos alertas. Ninguna combinación de paquetes RS mejoró la relación entre alertas y TP. Curiosamente, FortiGate no detectó estos ataques en los tres primeros días, mientras que Palo Alto activó solo una firma, generando tres alertas coincidentes con ataques verificados manualmente.

Si extendemos los resultados de al periodo completo de 30 días, nos encontramos que Snort activa entre 1 y 29 firmas que son TP (para RS1 y RS4), que dan lugar a entre 3 y 22 alertas diarias respectivamente. El resto de firmas activadas (FP) concentran el 99,9 % de las alertas generadas. Podemos comprobar que esta sensibilidad de Snort para la detección viene acompañada de un número muy elevado de falsos positivos, tanto mayor cuanto más elevado es el número de firmas del paquete o RS empleado. En el caso de FortiGate, las tres firmas que han saltado ninguna ha sido considerada como TP, mientras que, en el caso de Palo Alto, de las 6

firmas activadas 4 han sido TP.

V. Discusión de los resultados y limitaciones

Los resultados muestran que Snort ofrece una detección más exhaustiva, pero con una elevada tasa de falsos positivos, mientras que las soluciones comerciales detectan menos amenazas con mayor precisión. Esto sugiere que su equilibrio en la generación de alertas está diseñado para minimizar falsos positivos y reducir la carga de validación de incidentes, aunque ciertos ataques pueden pasar inadvertidos. No obstante, aunque los SIDS comerciales generan muchas menos alertas que Snort, un volumen cercano a 1000 alertas diarias sigue siendo excesivo para la mayoría de instalaciones residenciales.

Los resultados parecen confirmar que la mayoría de los ataques reales generan pocas alertas, lo que respalda la hipótesis de que las firmas que generan un número excesivo de alertas suelen ser Falsos Positivos. En la mayoría de los casos, estos FP se deben a errores de configuración o paquetes mal formados. Por ejemplo, la firma "UPnP.SSDP.M.Search.Anomaly" de Fortinet, clasificada como FP, generó 3614 alertas de severidad baja debido a dispositivos que generan solicitudes UPnP anómalas (mal formateadas o excesivas). Aunque este firma podría ser indicativa de malware o ataques DDoS, en entornos Smart Home, donde abundan dispositivos de bajo coste con configuraciones deficientes, es común que generen alertas sin representar una amenaza real. Esto resalta la necesidad de ajustar las configuraciones de los cortafuegos comerciales para estos entornos, ya que están optimizados para redes empresariales con énfasis en amenazas externas.

Por otro lado, si consideramos solo las firmas que generan verdaderos positivos (TP) en Snort, la tasa de alertas diarias se reduce a un nivel manejable (entre 3 y 22, según el conjunto de firmas). Además, la mayoría de los FP provienen de unas pocas firmas que no deberían aplicar para el entorno evaluado, lo que sugiere que un ajuste adecuado es clave para la viabilidad de Snort en redes residenciales. En este sentido, el algoritmo propuesto en [4] para desactivar firmas que generan FP en tráfico HTTP podría extenderse a otros protocolos, facilitando el uso de Snort en entornos Smart Home. Esto mismo aplicaría a los productos comerciales, aunque habitualmente no permiten al usuario un ajuste directo de las firmas utilizadas.

V-A. Limitaciones del trabajo realizado

Debido al elevado volumen del dataset utilizado no ha sido posible realizar un análisis manual exhaustivo a fin de clasificar los paquetes como legítimos o de ataque. La ausencia de *Ground Truth* limita la generación de los resultados presentados. En el caso de los tres primeros días de captura se ha realizado una validación manual de todas las alertas obtenidas por Snort, inspeccionando los paquetes asociados a dichas alertas a fin de determinar si son TP. Sin embargo, esta inspección solo abarcó los paquetes asociados a dichas alertas, por lo que es posible que algunos ataques hayan pasado inadvertidos. En cualquier caso, cualquier omisión habría afectado por igual a los tres SIDS evaluados, por lo que no se espera un impacto significativo en la comparación entre ellos.

Otra aproximación adoptada fue considerar como verdaderos positivos (TP) todas las alertas asociadas a una firma previamente clasificada como TP. Aunque en ciertos escenarios esta suposición podría no ser completamente precisa, la validación manual de las alertas en los tres días analizados, junto con la proporción de alertas observadas, sugiere que el tráfico en entornos Smart Home presenta una alta periodicidad. Esto hace que la metodología empleada sea razonablemente adecuada para nuestro caso de estudio.

Además, los resultados obtenidos considera el uso de los equipos con las firmas y configuraciones por defecto. Posiblemente tanto Snort como los equipos comerciales se podrían configurarse mejor a entornos Smart Home mediante la eliminación de alertas generadas por firmas identificadas como falsos positivos (FP). Esto mejoraría su aplicabilidad en escenarios reales sin necesidad de recurrir a sistemas SIEM y correlacion de eventos.

A pesar de estas limitaciones, consideramos que los resultados reflejan las tendencias cualitativas del comportamiento de los SIDS evaluados y proporcionan información valiosa en un área con escasos estudios experimentales.

VI. Conclusiones y líneas de avance

El análisis comparativo evidencia que los Sistemas de Detección de Intrusos basados en firmas presentan limitaciones en entornos SmartHome, donde la alta tasa de alertas es inasumible para la mayoría de instalaciones residenciales. Con respecto a la comparación entre detectores de código abierto y comerciales, podemos concluir que no es acertado realizar una comparación directa sin considerar sus enfoques y objetivos, ya que operan con criterios distintos. Las soluciones comerciales como FortiGate y Palo Alto NGFW priorizan la reducción de falsos positivos, generando menos alertas pero con menor sensibilidad ante ciertos ataques. Además, estos dispositivos, diseñados como cortafuegos avanzados con capacidad de bloqueo (IPS), deben minimizar falsos positivos para evitar interrupciones en redes corporativas.

Se ha constatado que Snort, sistema de código abierto con una amplia base de firmas, ofrece una detección más exhaustiva, pero a costa de una elevada tasa de falsos positivos, lo que dificulta su implantación sin un proceso de filtrado y correlación de eventos. Estos resultados ponen de manifiesto la necesidad de desarrollar metodologías de ajuste y optimización de firmas específicas para entornos SmartHome, dado que las herramientas evaluadas operan con conjuntos de firmas diseñados para redes corporativas o industriales, no necesariamente adaptados a las características del tráfico doméstico. Asimismo, podemos concluir que la falta de un marco de comparación adecuado con nuevas métricas adaptadas (benchmarks) impide realizar una comparación objetiva del rendimiento de los diferentes SIDS en estos entornos.

Estos resultados, aunque preliminares, son relevantes y el trabajo sigue en desarrollo. Como líneas futuras, se propone implementar mecanismos de filtrado para reducir falsos positivos en Snort y evaluar estrategias híbridas que combinen firmas con modelos de detección por anomalías. También se plantea supervisar manualmente el tráfico capturado durante 30 días para mejorar la clasificación de eventos. Finalmente, se explorará la integración de inteligencia artificial para adaptar dinámicamente las firmas según el tráfico y comportamiento de los dispositivos SmartHome.

AGRADECIMIENTOS

A las empresas Fortinet y Palo Alto Networks por facilitar el acceso a sus dispositivos, así como a sus ingenieros de soporte por su valiosa asistencia en los procesos de instalación, configuración y valoración de los resultados.

REFERENCIAS

[1] E. Castillo-Fernández, J. Diaz-Verdejo, R. Estepa Alonso, A. Estepa Alonso, Riesgos en la Smart Home: estudio experimental", en *Actas de las VIII Jornadas Nacionales de Investigación en Ciberseguridad (JNIC23)*, 375-382, 2023.

[2] D. Bonaventura, S. Esposito, G. Bella, "The IoT Breaches Your Household Again", en *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024)*, pp. 475–482, 2024.

[3] G. Vardakis G, G. Hatzivasilis, E. Koutsaki, N. Papadakis, Review of Smart-Home Security Using the Internet of Things", Electronics, 13(16):3343, 2024.

[4] J. Díaz-Verdejo, F.J. Muñoz-Calle, A. Estepa Alonso, R. Estepa Alonso, G. Madinabeitia, .On the detection capabilities of signature-based intrusion detection systems in the context of web attacks..Applied Sciences 12.2: 852, 2022.

[5] Gajewski, Mariusz, et al., .Anomaly traffic detection and correlation in smart home automation IoT systems", Transactions on Emerging Telecommunications Technologies 33(6) p. e4053, 2022.

[6] Cisco Systems, Inc. Snort: Network Intrusion Detection and Prevention System, https://www.snort.org/ (Último acceso 15 febrero 2025).

[7] Fortinet, Inc. FortiGate: Next-Generation Firewall (NGFW), https://www.fortinet.com/products/next-generation-firewall (Último acceso 15 febrero 2025).

[8] Palo Alto Networks, Inc. Next-Generation Firewall, https://docs.paloaltonetworks.com/ngfw (Último acceso 15 febrero 2025).

[9] R. Heartfield, G. Loukas, A. Bezemskij and E. Panaousis, "Self-Configurable Cyber-Physical Intrusion Detection for Smart Homes Using Reinforcement Learning,.en IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1720-1735, 2021.

[10] F. Alsakran, G. Bendiab, S. Shiaeles, N. Kolokotronis, N. Ïntrusion detection systems for smart home iot devices: experimental comparison study", en *International symposium on security in computing and communication*, pp. 87-98, 2019.

[11] N. Elsayed, Z. Zaghloul, S. Azumah, C. Li, Ïntrusion detection system in smart home network using bidirectional LSTM and convolutional neural networks hybrid model", en 2021 IEEE international midwest symposium on circuits and systems (MWSCAS), pp. 55–58, 2021.

[12] D. Rani, N. Gill, P. Gulia, F. Arena, G. Pau, "Design of an intrusion detection model for IoT-enabled smart home", IEEE Access, 11, 52509-52526,2023.

[13] A. Lara, V. Mayor, R. Estepa, A. Estepa, J.E. Díaz-Verdejo, "Smart home anomaly-based IDS: Architecture proposal and case study", Internet of Things, vol. 22, pp. 100773, 2023.

[14] Fred Klassen, AppNeta. Tcpreplay: Pcap editing and replaying utilities, https://tcpreplay.appneta.com/ (Último acceso 15 febrero 2025).

[15] Cisco Talos. Ruleset Talos, https://www.snort.org/talos/ (último acceso 01 marzo 2025).

[16] Emerging Threats. Proofpoint Emerging Threats Rules, https://rules.emergingthreats.net/ (último acceso 01 marzo 2025).

AutoLoRaSeguro: Proceso de Registro Automático Seguro para Dispositivos LoRaWAN

Lucas Mallen, Julio A. Sanguesa, José Roldán-Gómez, Vicente Torres-Sanz, Francisco J. Martinez
Departamento de Ingeniería Informática e Ingeniería de Sistemas
Universidad de Zaragoza, España
{l.mallen, jsanguesa, jroldan, vtorres, f.martinez}@unizar.es

***Resumen*—El registro de dispositivos en sistemas LoRaWAN puede ser un proceso tedioso, sobre todo si se han de registrar una gran cantidad de dispositivos. El sistema *AutoLoRaConfig* permite mitigar este problema, ya que permite el registro de múltiples dispositivos de manera automática, pero el uso de dicho sistema presenta vulnerabilidades.**

Este artículo propone *AutoLoRaSeguro*, un protocolo especialmente diseñado para mejorar la seguridad en el proceso de registro automático de dispositivos en redes LoRaWAN. En concreto, *AutoLoRaSeguro* permite evitar que, al usar *AutoLoRaConfig*, los atacantes puedan acertar el *DevEUI invitado* e introducir dispositivos maliciosos en la red. Para ello, el protocolo propuesto implementa una técnica de ofuscación de datos de los mensajes de registro y añade una capa adicional de autenticación mediante un código de seguridad configurable, que permite determinar la estructura y el orden de los *DevEUI* dentro de los paquetes de registro, lo que dificulta exponencialmente el registro no autorizado.

Los resultados del rendimiento de *AutoLoRaSeguro* demuestran que este reduce la probabilidad de éxito de los atacantes. Además, se ha llevado a cabo un análisis comparativo con los cifrados AES y SPECK, en el que se demuestra que la ofuscación de *AutoLoRaSeguro* proporciona un enfoque de seguridad robusto y eficaz.

***Index Terms*—Ciberseguridad, LoRaWAN, IoT, registro automático, ofuscación de datos, autenticación.**

I. Introducción

La comunicación inalámbrica se ha convertido en un elemento indispensable, que facilita la conexión entre dispositivos sin la dependencia de cables. Este tipo de comunicación ha sido imprescindible para muchos ámbitos, uno de ellos es el Internet de las Cosas (IoT) [1]. Este se refiere al uso de tecnologías para la interconexión de dispositivos físicos que permiten recopilar información del entorno y se comunican entre ellos.

Una de las tecnologías más utilizadas en las comunicaciones IoT ha sido LoRa (Long Range), ya que ofrece una comunicación a largo alcance, con un bajo consumo de energía y hace uso de bandas abiertas. Por otra parte, LoRaWAN (Long Range Wide Area Network) [2] es un protocolo de red diseñado para gestionar comunicaciones inalámbricas en IoT. Funciona sobre la tecnología LoRa, permitiendo la conexión de dispositivos a través de gateways que redirigen los datos a un servidor central.

Uno de los problemas que presenta este protocolo es la autenticación de los dispositivos en la red. En este proceso, es necesario manejar de forma precisa ciertos valores únicos, como el *DevEUI*, el *AppEUI* y la *AppKey*. Estos valores deben configurarse, tanto en el dispositivo como en el servidor, asegurando que coincidan en ambos para permitir la comunicación.

Para cada uno de los dispositivos que se quieran conectar a la red habrá que realizar este proceso, obteniendo el DevEUI y la AppKey del dispositivo y añadiéndola en el servidor. Esto provoca que registrar múltiples dispositivos se convierta en un proceso lento y propenso a errores. Para solucionar este problema, Torres-Sanz et al. [3] presentaron *AutoLoRaConfig*, un sistema diseñado para automatizar el proceso de registro y configuración de dispositivos en redes LoRaWAN.

AutoLoRaConfig automatiza el registro mediante el uso de reglas predefinidas que especifican el conjunto de dispositivos que tienen permitido registrarse en la red. Este proceso automatizado no solo acelera el registro, sino que también reduce el riesgo de errores, ya que elimina la necesidad de configurar manualmente cada dispositivo de manera individual. Sin embargo, la autenticación múltiple de distintos dispositivos puede presentar problemas potenciales de seguridad y provocar vulnerabilidades en la red.

La principal vulnerabilidad identificada es que permite la autenticación de cualquier dispositivo que haga uso de un *DevEUI invitado*, o lo que es lo mismo, que su *DevEUI* esté incluido dentro de los rangos especificados en las reglas de acceso. Esto significa que, si un atacante logra descubrir un valor válido de *DevEUI*, podría registrar múltiples dispositivos no autorizados, poniendo en peligro la integridad de la red.

Este artículo propone *AutoLoRaSeguro*, un protocolo especialmente diseñado para la mejora de la seguridad en el proceso de autenticación y registro automático de dispositivos LoRaWAN, protegiendo el sistema contra posibles dispositivos maliciosos. Más en detalle, *AutoLoRaSeguro* combina el uso de un código de seguridad junto con técnicas de ofuscación de datos, lo que asegura que la información transmitida se mantenga protegida, e impide que un dispositivo que no tenga el código se pueda registrar. Pese a que un atacante logre interceptar los paquetes, los datos serán incomprensibles, garantizando así un paso seguro de la información y reforzando la seguridad del sistema.

El artículo se estructura de la siguiente manera: la Sección II muestra las potenciales amenazas de seguridad que presenta *AutoLoRaConfig*. En la Sección III se explican algunas soluciones aplicadas en otras redes, así como los problemas de seguridad en LoRa. En la Sección IV se explica en detalle la solución propuesta. Los resultados obtenidos se presentan en la Sección V. Finalmente, la Sección VI muestra las conclusiones obtenidas.

II. Modelo de amenazas

En esta sección, se enumeran las asunciones que suponemos relativas a los atacantes, así como las acciones que estos

pueden llevar a cabo para comprometer el sistema.

Una vulnerabilidad potencial de la aplicación *AutoLoRaConfig* surge debido a que, para que un usuario del sistema realice el registro de dispositivos, se requieren dos variables: el *DevEUI invitado* y la *AppKey* .

El *DevEUI invitado* debe estar incluido en el conjunto de los previamente definidos, de manera secuencial, en cada una de las reglas. Estos se habilitan cuando se crea una nueva regla y se deshabilitan cuando se elimina. Esto implica que, en el caso de que un atacante conociera uno de los *DevEUI invitados*, podría obtener todos los pertenecientes a la regla en la que está contenido.

Por otro lado, de acuerdo al sistema *AutoLoRaConfig*, la *AppKey* requerida para realizar el registro será común a todas las reglas, y será necesario que los usuarios la conozcan para registrar nuevos dispositivos. Esto facilita su obtención por parte de los atacantes, ya que los usuarios deben transmitirla en alguna ocasión.

En este punto, se considera que los atacantes pueden obtener la *AppKey* de la aplicación *AutoLoRaConfig*, ya sea porque la han robado de alguno de los usuarios, o porque han sido usuarios del sistema con anterioridad. Esto les permite agregar dispositivos simplemente obteniendo uno de los *DevEUI invitado*.

Además, se considera que el atacante tiene la capacidad de usar la *AppKey* para decodificar el contenido de los paquetes de registro que se envían, lo que le permite ver los *DevEUI* reales de los dispositivos que se están registrando en la red.

Teniendo en cuenta todo lo anterior, este artículo presenta una capa de seguridad para evitar que algún atacante registre dispositivos maliciosos en la red, incluso en el caso de que adivine uno de los *DevEUI invitado*.

III. Estado del Arte

Los sistemas que permiten el registro automático en una red se enfrentan a múltiples desafíos de seguridad, especialmente en el contexto de redes heterogéneas e IoT. Estos sistemas pueden verse comprometidos por vulnerabilidades como configuraciones incorrectas, problemas de autenticación y protocolos poco robustos. La diversidad de dispositivos y la naturaleza dinámica de estos entornos amplifican los riesgos, exponiendo las redes a ataques de interceptación o manipulación de datos. Las amenazas más comunes son la interceptación de mensajes, la suplantación de identidad y la explotación de fallos en la autenticación, comprometiendo la privacidad y permitiendo accesos no autorizados.

La literatura aborda diversos mecanismos de autenticación para el registro de dispositivos. Saqib et al. [4] realizan una revisión sistemática de la literatura (SLR) sobre problemas de seguridad en IoT desde la perspectiva de la autenticación. Destacan la importancia de la autenticación en redes de recursos limitados, como IoT, y examinan marcos de autenticación basados en identidad, criptografía, tokens, hardware, procedimientos y contexto. Evalúan estos marcos según su coste computacional, sobrecarga de comunicación y consumo de energía.

Nandy et al. [5] revisan los mecanismos de autenticación en IoT, enfocándose en los desafíos de seguridad y las soluciones propuestas. Más en detalle, identifican métodos como contraseñas, tokens, biometría y criptografía, analizando sus fortalezas y debilidades. Además, profundizan en ataques a la autenticación en IoT (suplantación, intermediarios, denegación de servicio) y analizan técnicas de evaluación como tiempo de respuesta y consumo de memoria, proporcionando una guía para la implementación de protocolos más seguros.

Ingles-Sanchez et al. [6] abordan el desafío de la incorporación segura de dispositivos IoT, centrándose en el proceso de arranque. En concreto, el estudio evalúa la funcionalidad del método EAP-NOOB (EAP-Nimble out-of-band) para el envío de los paquetes en redes LoRaWAN. Ante las limitaciones de tamaño de mensaje en LoRaWAN, los autores proponen el uso de LO-CoAP-EAP como capa inferior de EAP y la representación binaria concisa de objetos (CBOR) para reducir el tamaño de los mensajes intercambiados.

En lo relativo a la ofuscación de datos, utilizada fundamentalmente para evitar la interceptación de mensajes, podemos encontrar diferentes propuestas. Hosseinzadeh et al. [7] proponen la ofuscación de protocolos de comunicación en IoT para ocultar patrones, dificultando el espionaje y la manipulación. Los autores argumentan que, la ofuscación combinada con la diversificación de software, mejora la seguridad, al introducir variabilidad y dificultando ataques a gran escala.

Hessler et al. [8] implementan técnicas de ofuscación mediante la codificación de la red, cifrando el vector de coeficientes que indica cómo se combinaron los paquetes originales, en lugar de cifrar los paquetes completos. Esto permite a los nodos intermedios manipular paquetes codificados sin revelar su composición.

Por su parte, Ghourab y Azab [9] proponen un mecanismo de defensa de objetivo móvil (MtD) para comunicaciones inalámbricas. En concreto, inyectan datos falsos y saltan entre frecuencias para ofuscar patrones de transmisión, dificultando el espionaje, incluso para atacantes con múltiples puntos de acceso.

Cabe destacar que, el problema de seguridad específico que presenta *AutoLoRaConfig*, donde los atacantes pueden registrar dispositivos maliciosos conociendo un *DevEUI invitado*, no se puede resolver utilizando alguna de estas técnicas propuestas, ya que estas soluciones no abordan la combinación particular de registro automático basado en rangos de *DevEUIs invitados* y la vulnerabilidad inherente a ese enfoque. Teniendo en cuenta lo anterior, este artículo propone un protocolo que combina mecanismos de autenticación robustos con ofuscación de datos durante el envío, protegiendo la información del paquete sin necesidad de encriptación completa.

IV. AutoLoRaSeguro: Un Protocolo de Seguridad para el Sistema AutoLoRaConfig

IV-A. Arquitectura de AutoLoRaConfig

AutoLoRaConfig es un sistema que permite el registro automático de múltiples dispositivos en una red LoRaWAN mediante OTAA, agilizando significativamente el proceso y mitigando los posibles errores que se pueden cometer durante el registro manual de los dispositivos.

Este sistema se basa en la definición de reglas, las cuales se utilizan para controlar el acceso de los dispositivos. Cada una de estas reglas contiene un rango de *DevEUI invitados*, que son los valores de los *DevEUIs* a los que se les permitirá el

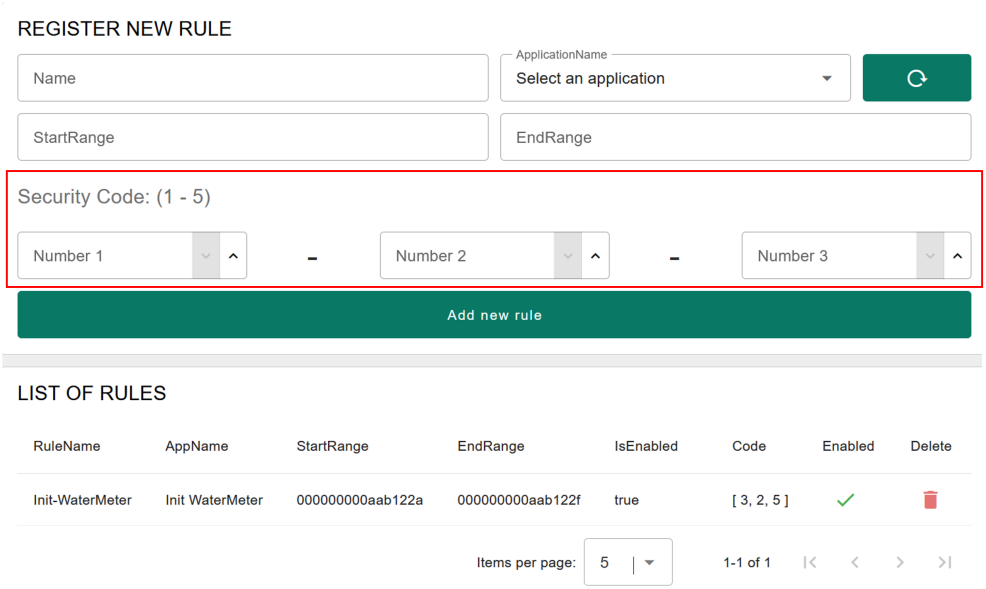


Figura 1: Interfaz de administración de AutoLoRaSeguro.

registro en el sistema, de manera que, cuando un dispositivo quiera registrarse deberá utilizar un *DevEUI invitado* válido, es decir, que esté contenido en el rango de una de las reglas. Además, cada una de las reglas contiene una *AppKey*, que representa el *id* de la aplicación a la que se va a conectar el dispositivo, una vez registrado.

Dado que cuando un dispositivo hace la conexión con el *DevEUI invitado* se añadirá su *DevEUI* real a la aplicación que esté establecida en la regla, las reglas no podrán contener rangos de *DevEUI* comunes entre ellas.

Al utilizar *AutoLoRaConfig*, el usuario únicamente debe crear la regla y configurar los dispositivos con alguno de los *DevEUIs* presentes dentro del rango y una *AppKey* de invitado común para todos, de manera que el sistema realice el registro de todos los dispositivos de forma automática.

Este sistema permite que un usuario pueda registrar múltiples dispositivos en la red LoRaWAN de manera rápida y sencilla, tardando un promedio de 45.1 segundos en realizar el registro de un dispositivo, frente a unos 88 segundos que se tardaría en hacerlo de manera manual, y un promedio de 55 segundos en registrar 100 dispositivos, frente a unos 7513 segundos que se tardaría en hacerlo de manera manual [3]. Gracias a *AutoLoRaConfig*, se mejora no solo la velocidad del registro de una gran cantidad de dispositivos en una red, sino que también se reduce la posibilidad de cometer fallos humanos durante el proceso de registro.

IV-B. Protocolo AutoLoRaSeguro

El sistema de *AutoLoRaConfig* plantea una forma de registrar múltiples dispositivos en una red LoRa de manera automática, acelerando el proceso de registro y minimizando en gran medida el error humano. Pero, como ya se ha mencionado, este sistema lleva consigo un inconveniente relacionado con la seguridad que permiten a atacantes introducir dispositivos maliciosos en la red únicamente acertando en *DevEUI invitado*.

Para solucionar este problema, en este artículo se diseña e implementa un nuevo protocolo para el registro de dispositivos en la red LoRaWAN, llamado *AutoLoRaSeguro*, el cual utiliza una técnica de ofuscación de los datos en los mensajes de registro que se envían por la red. Este protocolo añade al registro automático un código, que proporciona un mecanismo de seguridad adicional y permite evitar el registro no autorizado de dispositivos, incluso en caso de que los atacantes tengan acceso al *DevEUI*.

Para ello, se ha modificado la parte del servidor que permite añadir nuevas reglas (ver Figura 1). Concretamente se ha añadido un nuevo campo al formulario de *AutoLoRaConfig*, que permite introducir una secuencia de seguridad de tres valores (entre el uno y el cinco), que representa el código que enviarán los dispositivos durante el proceso de registro.

En el sistema anterior, el mensaje de registro enviado a través de la red utilizaba un *DevEUI invitado*, configurado previamente en las reglas, e incluía en el *payload* el *DevEUI* real y la *AppKey*, ambos necesarios para añadir un nuevo dispositivo a la red. El protocolo propuesto en este artículo requiere también transmitir esta información, pero introduce una capa adicional de seguridad mediante la ofuscación de los mensajes enviados.

El proceso de registro completo implica la transmisión de varios paquetes, como se muestra en la Figura 2. Cada paquete incluye un identificador único que indica su posición en la secuencia de envío, seguido de la *AppKey*, que debe ser consistente en todos los mensajes de registro enviados por el dispositivo. Además, cada paquete contiene un *DevEUI*, el cual puede ser el *DevEUI* real del dispositivo o un *DevEUI* falso, en función del número de paquete que sea. Cuando los paquetes llegan al servidor, este verifica si la secuencia de seguridad se cumple para dar de alta el nuevo dispositivo.

La forma en la que funciona el código de seguridad es que cada uno de los valores que se establecen en el nuevo campo del servidor hacen referencia a la posición del paquete en la que se ubica *DevEUI* real, de manera que si se establece el

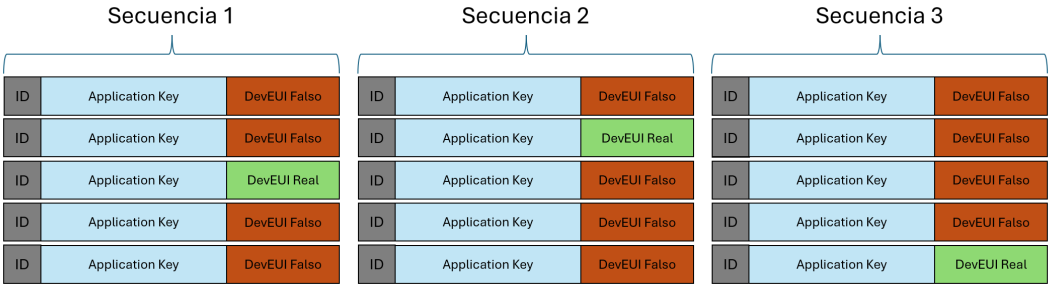


Figura 2: Payload de los paquetes de registro.

código de seguridad con los valores 3, 2 y 5, el servidor espera a que le lleguen todos los paquetes que envía el dispositivo y los combina de cinco en cinco (que es el número de *DevEUIs* que se ha establecido por cada secuencia). Tras eso, se comprueba que el tercer *DevEUI* de la primera secuencia coincida con el segundo *DevEUI* de la segunda secuencia y con el quinto *DevEUI* de la tercera secuencia. En el caso de que coincidan los *DevEUI*, se confirma que el dispositivo es auténtico y se permite su registro.

Además, se hacen otro tipo de comprobaciones para asegurar el registro, como comprobar que el tamaño de los paquetes es el mismo, que se han enviado un número valido de secuencias y de *DevEUIs*, que no hay *DevEUIs* repetidos en una misma secuencia, y que no se hayan enviado paquetes con el mismo *id*.

Al utilizar este protocolo, los atacantes no solo deben acertar el *DevEUI invitado*, sino que también deben acertar el código de seguridad para el envío de los paquetes, indicado en la regla. Este sistema no solo amplía la seguridad, dificultando el registro de dispositivos maliciosos, sino que además ofusca el *DevEUI* frente a posibles intercepciones de paquetes, haciendo que, en caso de que los atacantes obtengan el paquete y lo decodifiquen, no sepan si el valor del *DevEUI* que contiene es el que se está añadiendo a la red o es uno falso.

La Figura 3 muestra cómo se produce el proceso de registro en *AutoLoRaSeguro*, desde que se inicia el dispositivo hasta que se confirma el registro de este. Como se observa, *AutoLoRaSeguro* cuenta con tres grandes componentes: el dispositivo, el servidor ChirpStack y la aplicación *AutoLoRaSeguro*.

Cuando se inicia el dispositivo lo primero que hace es comprobar si está registrado o no; en caso de que esté registrado, se obtiene el *DevEUI* del chip, que es un código único para cada dispositivo físico LoRa, y la *AppKey*, que se utiliza para la conexión, y se envía un *JoinRequest* para poder iniciar el envío de datos. En caso de que el dispositivo no esté registrado, se utilizan los valores de invitado de *DevEUI* y *AppKey*, que son los valores establecidos por defecto, y se envía un *JoinRequest*. Este mensaje llega al servidor ChirpStack, que comprueba si el *DevEUI* existe, es decir, que está en el rango de una de las reglas activas, y si es así, le envía un *JoinAccept* para que prosiga la comunicación.

Una vez el dispositivo recibe el *JoinAccept*, obtiene el *DevEUI* del chip y genera una *AppKey* y una lista de *DevEUI* falsos para crear los mensajes del registro. Como se ha visto anteriormente, cada uno de estos paquetes contiene: (i) un *id*, que es único para cada paquete y representa la posición en la que se envía, (ii) la *AppKey*, que ha de ser igual en todos los paquetes que se envían, y (iii) un *DevEUI*, que puede ser falso o verdadero en función del número del paquete y de acuerdo al código de seguridad.

Estos paquetes son enviados a la aplicación, que los recibe y realiza una serie de comprobaciones para confirmar que son válidos:

- Primero se comprueba que el *DevEUI invitado* con el que se accede pertenece a una regla.
- En caso afirmativo, se confirma que el paquete es válido, y se comprueba que el tamaño de este es adecuado, que la *AppKey* coincide con los demás paquetes que se transmiten, y que no se han enviado paquetes con el mismo *id*.
- Tras haber obtenido todos los paquetes, se preparan los *DevEUIs* que se van a enviar de acuerdo a la secuencia del código de seguridad. Para ello, se utiliza el número de los paquetes, de manera que, si cada secuencia contiene cinco números, se combinan los *DevEUIs* de los paquetes del 1 al 5, del 6 al 10 y del 11 al 15.
- Una vez se han creado las secuencias del código de seguridad se comprueban si estas son válidas. Para ello, se comprueba que el número de *DevEUIs* por secuencia sea adecuado, en caso de que no estuvieran todos los *DevEUIs* se invalidaría dicha secuencia. Además, se comprueba que no existan *DevEUIs* repetidos en las secuencias y que no se transmitan paquetes con un número *DevEUIs* superior al esperado, si ocurre alguno de estos casos se cancela el registro.
- A continuación, se comprueba que el código de seguridad sea el correcto, comprobando los *DevEUI* de las posiciones indicadas en la regla.
- En caso de que el código de seguridad sea el correcto, se acepta el registro y se envía una petición a ChirpStack para añadir el nuevo dispositivo a la red.
- En el caso contrario, la aplicación bloqueará al dispositivo, impidiendo que pruebe con otras combinaciones.

Una vez finaliza el registro, el dispositivo ya puede utilizar su *DevEUI* propio para comenzar con el envío de información.

IV-C. Control de la Pérdida de Paquetes

Este protocolo funciona de forma óptima en situaciones teóricas, pero en la práctica, cabe la posibilidad de que se pierdan algunos de los paquetes enviados usando comunicaciones

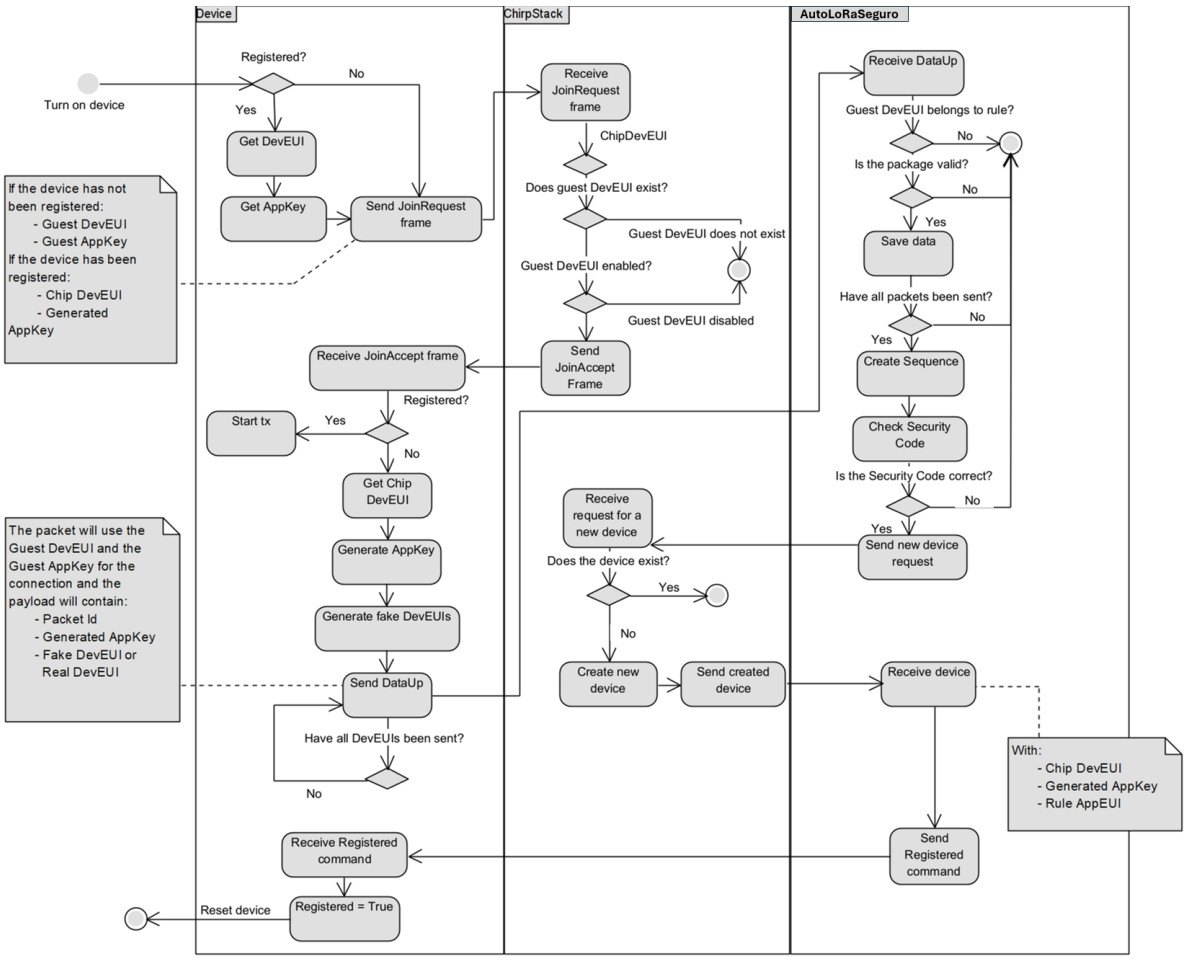


Figura 3: Diagrama de flujo de la estrategia de la seguridad para el registro automático de dispositivos LoRaWAN.

LoRa. Esto provoca que el proceso de registro automático de dispositivos utilizando *AutoLoRaSeguro* pueda no ser viable.

El comportamiento inherente del protocolo es muy estricto, provocando que sea necesario que el servidor reciba todas las secuencias con todos los *DevEUIs* en cada una de ellas, es por ello que la pérdida de un único paquete pueda ocasionar la cancelación del proceso de registro automático.

Para solucionar este problema, se ha optado por relajar ligeramente los requerimientos, de manera que el servidor únicamente necesite dos de las tres secuencias completas. De esta manera reducimos la probabilidad de que el proceso de registro se vea afectado por problemas en las comunicaciones, y que sea tolerante a fallos de comunicación, ya que aunque se pierdan algunos de los paquetes del registro, siempre y cuando estos sean de la misma secuencia, se podrá continuar con el proceso de registro de los dispositivos.

No obstante, cabe destacar que esta ligera modificación, a pesar de solucionar en gran medida el problema de la pérdida de paquetes, aumenta también la vulnerabilidad del sistema, ya que ahora los atacantes únicamente necesitan acertar dos de las tres secuencias del código de seguridad para poder acceder a la red. Pese a eso, y dado que la probabilidad de que se pierdan paquetes es mucho mayor, es razonable relajar los requisitos del protocolo, de manera que se obtenga un sistema de registro automático seguro, pero también funcional.

V. Resultados

V-A. *Análisis de la Probabilidad de Éxito del Atacante*

Para validar el funcionamiento de *AutoLoRaSeguro*, en esta subsección se muestra el cálculo de la probabilidad teórica de éxito, que tienen los atacantes al intentar introducir un dispositivo malicioso en la red. Para ello, se simulan varios escenarios en los que se cambia el número de *DevEUI* falsos que se envían por paquete y el número de paquetes que valida el servidor para aceptar el registro del dispositivo. Para calcular la proclividad que tienen los atacantes de registrarse, se ha utilizado la siguiente función:

$$P = \frac{\text{Combinaciones Favorables}}{\text{Combinaciones Posibles}}$$

En la que $Combinaciones\ Favorables$ hace referencia al número de casos en los que el *DevEUI* real está en la posición correcta dentro de la secuencia de paquetes que se va a validar, mientras que $Combinaciones\ Posibles$ considera todas las combinaciones de *DevEUI* falsos, junto con el *DevEUI* real.

La Tabla I muestra la probabilidad de éxito que tienen los atacantes a la hora de intentar acertar el código de seguridad para registrar un dispositivo en la red. Como se puede observar, la probabilidad disminuye a medida que va aumentando el número de *DevEUIs* que se utilizan en cada

Tabla I: Probabilidad de adivinar la combinación, basada en el número de *DevEUIs* por secuencia y el número de secuencias.

DevEUIs/ Secuencia	Número de Secuencias	Probabilidad con AutoLoRaSeguro	Probabilidad con AutoLoRaSeguro y Pérdida de Paquetes
3	2	11.11 %	33.33 %
	3	3.70 %	11.11 %
	4	1.23 %	3.70 %
	5	0.41 %	1.23 %
4	2	6.25 %	25.00 %
	3	1.56 %	6.25 %
	4	0.39 %	1.56 %
	5	0.10 %	0.39 %
5	2	4.00 %	20.00 %
	3	0.80 %	4.00 %
	4	0.16 %	0.80 %
	5	0.03 %	0.16 %
6	2	2.78 %	16.66 %
	3	0.46 %	2.78 %
	4	0.08 %	0.46 %
	5	0.01 %	0.08 %
7	2	2.04 %	14.28 %
	3	0.29 %	2.04 %
	4	0.04 %	0.29 %
	5	0.01 %	0.04 %

secuencia y a medida que se aumenta el número de secuencias que se emplean para validar el código.

Por una parte, si se utilizan 3 *DevEUIs* por secuencia y se tienen únicamente 2 secuencias, la probabilidad de que el atacante acierte es del 11.11 %. Sin embargo, al aumentar el número de *DevEUIs* a 4, pero manteniendo el número de secuencias constante, la probabilidad disminuye a 6.25 %. Este cambio muestra cómo un aumento en la complejidad del número de *DevEUIs* contribuye a dificultar el ataque, aunque no de manera drástica.

Por otro lado, si se decide mantener 3 *DevEUIs* y aumentar el número de secuencias de 2 a 3, la probabilidad se reduce a 3.70 %. Aquí se observa una mejora más significativa, ya que incrementar las secuencias implica una mayor complejidad exponencial. No obstante, el verdadero impacto se aprecia al incrementar ambos parámetros. Si se utilizan 4 *DevEUIs* y 3 secuencias, la probabilidad se reduce significativamente, alcanzando apenas el 1.56 %, lo que implica que se ha disminuido casi diez veces la probabilidad de que se acierte.

En esta misma tabla se puede observar cómo afecta el uso del protocolo mejorado para mitigar el efecto de la pérdida de paquetes. Al hacer uso de este se puede ver como los atacantes tienen una mayor probabilidad de acertar el código de seguridad que en los casos en los que no se utiliza el protocolo, variando desde un aumento del 11.11 % al 33.33 % (lo que es tres veces mayor) en el caso de utilizar 2 secuencias con 3 *DevEUIs* por secuencia hasta un aumento del 0.04 % al 0.29 % (lo que es 7.25 veces mayor) en el caso de utilizar 4 secuencias con 7 *DevEUIs* por secuencia.

Teniendo en cuenta los resultados obtenidos, se observa que se puede aumentar en gran medida la seguridad del protocolo, pero es importante considerar otros factores. Por ejemplo, aumentar el número de secuencias necesarias también provoca un aumento en el tiempo de registro de los dispositivos que queremos registrar. Además, al aumentar el número de paquetes que se envían, se aumenta el número de paquetes que se pueden perder, lo que también aumentaría el tiempo de registro de un dispositivo.

V-B. Análisis de la Pérdida de Paquetes

En esta subsección se analiza con más detenimiento el efecto que produce la pérdida de paquetes durante el proceso de registro de dispositivos, de acuerdo a los datos obtenidos. Para ello, se han realizado una serie de pruebas de registro donde se han modificado los data rates de envío, haciendo uso de 16 secuencias, con cinco *DevEUI* por secuencia y un total de 80 paquetes enviados.

Se puede observar que, con data rate 4 y 5 no se pierde ninguno de los paquetes. Además, con data rate 3 únicamente se pierde uno de los paquetes. Por otra parte, al utilizar data rate 2 se pierden tres paquetes, por lo que se ha perdido casi un 5 % de ellos. Con el uso de data rate 1 se pierden 5 paquetes, un 6.25 % de los paquetes. Por último con data rate cero se pierden 8 paquetes, siendo el que más impacto tiene en la de pérdida de paquetes con un 13.75 %.

V-C. Comparación con Algoritmos de Cifrado

Con el fin de contextualizar la eficiencia de la técnica de ofuscación implementada en *AutoLoRaSeguro*, se ha llevado a cabo un análisis comparativo del rendimiento del protocolo propuesto frente a dos algoritmos de cifrado simétrico comúnmente utilizados en entornos IoT. El propósito de esta comparación no es sugerir que la ofuscación deba reemplazar el cifrado en todos los escenarios posibles, sino demostrar su viabilidad como una capa de seguridad seguridad ligera y orientada a reducir el error humano, particularmente adecuada para el proceso de registro inicial automático en redes LoRaWAN.

Para las pruebas se ha utilizado un dispositivo Heltec Wifi LoRa 32 V3, que cuenta con un microcontrolador ESP32-S3FN8 y un módulo LoRa SX1262 que le permite comunicarse al servidor ChirpStack.

Se pretende comparar el rendimiento de *AutoLoRaSeguro* con el de los algoritmos de cifrado simétrico AES-128 (Advanced Encryption Standard) [10] y SPECK [11]. AES-128, un estándar ampliamente adoptado, es conocido por su equilibrio entre seguridad y rendimiento; en este análisis, se emplea en modo CBC (Cipher Block Chaining) con un vector de inicialización (IV) aleatorio. SPECK, por otro lado, es un algoritmo de cifrado de bloques ligero diseñado por la NSA, optimizado para microcontroladores y dispositivos con recursos limitados; en este análisis, se evalúa la variante SPECK-64/128 (bloque de 64 bits, clave de 128 bits).

Es importante destacar que tanto AES como SPECK requieren una gestión previa de claves que se realiza de manera manual. Esto implica el manejo de claves en formato hexadecimal de 16 bytes, lo que añade complejidad al proceso de registro y aumenta la probabilidad de errores humanos en entornos no gestionados automáticamente. En contraste, *AutoLoRaSeguro* permite codificar la autenticación utilizando simples secuencias numéricas entre 1 y 5, mucho más manejables y adecuadas para usuarios no expertos o procesos de configuración masiva.

En primer lugar, se ha evaluado el tiempo requerido por cada uno de los mecanismos para preparar el paquete antes de ser enviado a través de la red. La Figura 4 muestra los resultados del análisis de tiempo de ejecución para las diferentes opciones evaluadas. Se observa una gran variación

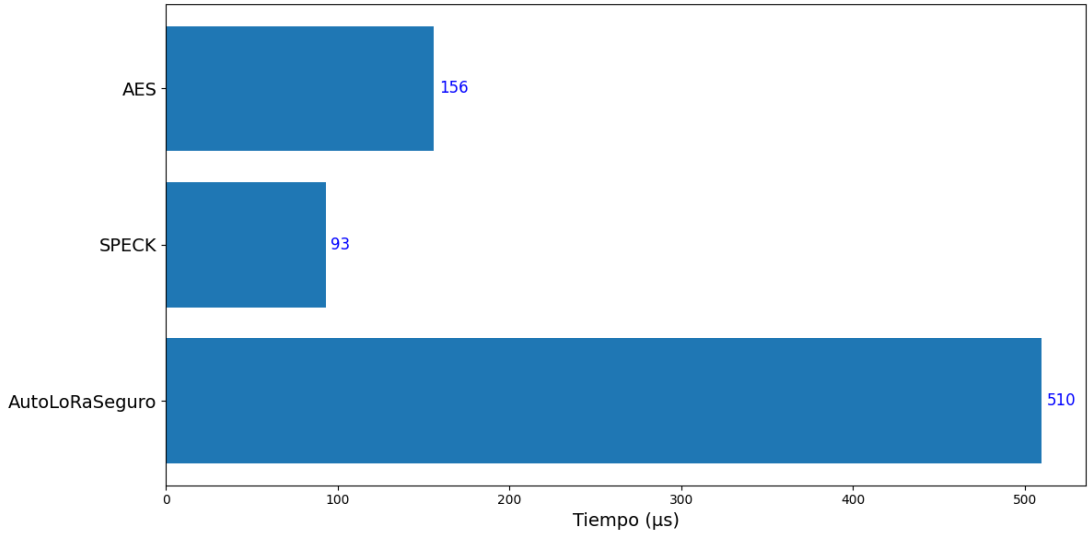


Figura 4: Comparación de tiempos de creación de paquetes de registro para diferentes técnicas.

en el rendimiento entre las dos opciones que usan métodos de cifrado puro (AES y SPECK, respectivamente) y las la opción que incorpora el protocolo *AutoLoRaSeguro*.

En los casos en los que AES y SPECK se aplican de forma aislada, los tiempos registrados fueron de 156 µs y 93 µs, respectivamente. Estos valores corresponden al tiempo necesario para el cifrado de un único paquete y con un tamaño menor, ya que no se incluye la sobrecarga de información asociada al protocolo *AutoLoRaSeguro*.

En el caso del uso de *AutoLoRaSeguro* se ha utilizado la configuración de tres secuencias de cinco *DevEUI* cada una. El resultado que se ha obtenido ha sido de únicamente 34 µs por paquete, pero al requerir de 15 paquetes para enviar todas las secuencias, el tiempo total requerido por la opción de *AutoLoRaSeguro* fue de 510 µs.

Por otra parte, se pretende conocer cómo afecta el uso de cada uno de los mecanismos en la carga de la red al registrar nuevos dispositivos. Para este propósito, se han tenido en cuenta las mismas consideraciones que en la prueba anterior y, adicionalmente, se ha analizado cómo afecta el data rate de LoRA al rendimiento de las distintas opciones.

Para complementar el análisis, la Tabla II muestra el Tiempo en el Aire (ToA) estimado para los mecanismos evaluados, en función del Spreading Factor (SF). Se observa que tanto AES como SPECK, al requerir únicamente un paquete de 48 bytes, mantienen un ToA significativamente inferior en todos los niveles de SF. Por ejemplo, a SF7, el ToA es de solo 112.9 ms, mientras que a SF12 se incrementa a 2793.5 ms.

En contraste, la implementación de *AutoLoRaSeguro* necesita 15 paquetes de 49 bytes, lo que multiplica el ToA total. A SF7, alcanza 1770 ms, y se eleva hasta 41902.5 ms en SF12. Este impacto está directamente relacionado con la fragmentación del mensaje de registro en múltiples paquetes, necesaria para transmitir todas las combinaciones codificadas.

Es importante destacar que, si bien la implementación de *AutoLoRaSeguro* conlleva un incremento notable en el ToA, hay que tener en cuenta que esto se circunscribe al procedimiento de registro inicial del dispositivo, el cual únicamente se va a realizar una vez en su ciclo de vida. El resto de mensajes

Tabla II: Tiempo en el Aire (ToA) de los mecanismos evaluados según el Spreading Factor.

Mecanismo	Bytes/Paquete	Paquetes	SF	ToA (ms)
AES	48	1	7	112.9
			8	205.3
			9	369.7
			10	698.4
			11	1478.7
			12	2793.5
SPECK	48	1	7	112.9
			8	205.3
			9	369.7
			10	698.4
			11	1478.7
			12	2793.5
AutoLoRaSeguro	49	15	7	1770.0
			8	3079.5
			9	5545.5
			10	10476.0
			11	22180.5
			12	41902.5

enviados, una vez completado el registro, transcurren sin esa sobrecarga adicional. Dada la naturaleza de las redes LoRaWAN, diseñadas para operaciones largas que abarcan años, este tiempo de registro inicial representa una fracción insignificante de la vida operativa del dispositivo.

VI. Conclusiones

En este trabajo se ha presentado *AutoLoRaSeguro*, un protocolo de seguridad diseñado para mejorar el proceso de registro automático de dispositivos en redes LoRaWAN, teniendo en cuenta las vulnerabilidades identificadas en el sistema *AutoLoRaConfig*. El protocolo propuesto implementa una estrategia de ofuscación de datos en los mensajes de registro y añade una capa adicional de seguridad mediante la introducción de un código de seguridad en forma de secuencia de posiciones de *DevEUI*.

Los resultados obtenidos demuestran que *AutoLoRaSeguro* reduce significativamente la probabilidad de que los atacantes puedan introducir dispositivos maliciosos en la red. Al aumentar el número de *DevEUIs* por secuencia y el número

de secuencias requeridas para el registro, la probabilidad de éxito de un ataque disminuye exponencialmente, alcanzando valores inferiores al 1 % en configuraciones sencillas.

Por otra parte, y dado que en redes LoRaWAN pueden perderse paquetes, se ha mejorado el protocolo para que pueda funcionar en caso de que exista pérdida de paquetes, a costa de sacrificar ligeramente la seguridad. Los resultados muestran que, una vez mejorado el protocolo, la pérdida de paquetes deja de ser un obstáculo, siempre que se reciban al menos dos secuencias completas, a costa de incrementar ligeramente la probabilidad de que un atacante pueda comprometer el sistema.

El análisis comparativo con los métodos de cifrado simétrico tradicionales (AES y SPECK) muestra que, si bien el tiempo de ejecución para preparar los paquetes de registro es mayor cuando se utiliza *AutoLoRaSeguro*, este aumento es asumible, y que vale la pena teniendo en cuenta el aumento de seguridad obtenido.

En conclusión, *AutoLoRaSeguro* aporta una solución efectiva para mejorar la seguridad en el registro automático de dispositivos LoRaWAN, combinando técnicas de ofuscación y autenticación que refuerzan la integridad y confidencialidad de la red sin comprometer significativamente su operatividad. No obstante, es importante reconocer que este protocolo añade cierta complejidad adicional en el proceso de registro y puede requerir ajustes en entornos con restricciones severas de ancho de banda o energía.

VII. Agradecimientos

Este trabajo ha sido parcialmente financiado por el Gobierno de Aragón y el Fondo Social Europeo 'Construyendo Europa desde Aragón' (Grupo de Investigación T40_23D), por el proyecto de I+D PID2021-122580NB-I00, financiado por MICIU/AEI/10.13039/501100011033 y 'FEDER. Una manera de hacer Europa', así como por fondos del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation), dentro del proyecto estratégico C077/23 del Instituto Nacional de Ciberseguridad (España).

Referencias

[1] S. Li, L. D. Xu, and S. Zhao, "The Internet of Things: a survey," *Information systems frontiers*, vol. 17, pp. 243–259, 2015.

[2] J. Haxhibeqiri, E. De Poorter, I. Moerman, and J. Hoebeke, "A survey of LoRaWAN for IoT: From technology to application," *Sensors*, vol. 18, no. 11, p. 3995, 2018.

[3] V. Torres-Sanz, J. A. Sanguesa, F. J. Martinez, P. Garrido, and C. T. Calafate, "An autoconfiguration strategy for very large scale long range wide area network deployments in smart cities," *IET Smart Cities*, vol. 7, no. 1, p. e12096, 2025.

[4] M. Saqib and A. H. Moon, "A Systematic Security Assessment and Review of Internet of Things in the Context of Authentication," *Computers and Security*, vol. 125, p. 103053, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740482200445X

[5] T. Nandy, M. Y. I. B. Idris, R. Md Noor, L. Mat Kiah, L. S. Lun, N. B. Annuar Juma'at, I. Ahmedy, N. Abdul Ghani, and S. Bhattacharyya, "Review on Security of Internet of Things Authentication Mechanism," *IEEE Access*, vol. 7, pp. 151 054–151 089, 2019.

[6] E. Ingles-Sanchez, D. Garcia-Carrillo, G. Z. Papadopoulos, N. Montavont, and A. F. Skarmeta Gómez, "Adaptation of EAP-NOOB Method for LoRaWAN with LO-CoAP-EAP and CBOR," in *2020 Global Internet of Things Summit (GIoTS)*, 2020, pp. 1–6.

[7] S. Hosseinzadeh, S. Hyrynsalmi, and V. Leppänen, "Chapter 14 - Obfuscation and diversification for securing the internet of things (IoT)," in *Internet of Things*, R. Buyya and A. Vahid Dastjerdi, Eds. Morgan Kaufmann, 2016, pp. 259–274. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128053959000149

[8] A. Hessler, T. Kakumaru, H. Perrey, and D. Westhoff, "Data obfuscation with network coding," *Computer Communications*, vol. 35, no. 1, pp. 48–61, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366410004779

[9] E. M. Ghourab and M. Azab, "Benign false-data injection as a moving-target defense to secure mobile wireless communications," *Ad Hoc Networks*, vol. 102, p. 102064, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570870519306304

[10] V. Rijmen and J. Daemen, "Advanced encryption standard," *Proceedings of federal information processing standards publications, national institute of standards and technology*, vol. 19, p. 22, 2001.

[11] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers, "The SIMON and SPECK lightweight block ciphers," in *Proceedings of the 52nd Annual Design Automation Conference*, ser. DAC '15. New York, NY, USA: Association for Computing Machinery, 2015. [Online]. Available: https://doi.org/10.1145/2744769.2747946

QUILLAQUA: Hacia un marco sostenible para entornos hidropónicos seguros y eficientes en la Agricultura 4.0.

Diego F. Bustamante V., Luis E. Sánchez, David G. Rosado,
Antonio Santos-Olmo, Eduardo Fernández-Medina
GSyA Research Group,
University of Castilla-La Mancha, Ciudad Real, Spain
diegofernando.bustamante@alu.uclm.es,
{luise.sanchez, david.grosado, antonio.santosolmo, eduardo.fdezmedina}@uclm.es

***Resumen*—La transformación digital de la agricultura demanda un enfoque integral que combine inteligencia artificial, ciberseguridad y sostenibilidad. En este artículo presentamos QUILLAQUA, un marco que optimiza la gestión de cultivos hidropónicos mediante la integración de sensores, IoT y modelos de IA para detectar y ajustar parámetros clave en tiempo real. Además, incorpora protocolos de ciberseguridad que protegen los datos y garantizan la fiabilidad de la infraestructura agrícola frente a amenazas crecientes. QUILLAQUA también promueve la eficiencia en el uso del agua y los nutrientes, reduciendo la huella medioambiental y fortaleciendo la resiliencia ante el cambio climático. Su enfoque holístico abarca la mejora continua de técnicas de producción, la reducción de costes energéticos y la formación de personal, fomentando la adopción de prácticas responsables. Con ello, la propuesta avanza hacia una Agricultura 4.0 segura, sostenible y capaz de afrontar los desafíos globales en materia de seguridad alimentaria.**

***Index Terms*—Ciberseguridad, Agricultura 4.0, IA, Hidroponía, IoT, Sostenibilidad, Análisis de Riesgos.**

Tipo de contribución: *Investigación original*

I. Introducción

La digitalización de la agricultura ha dado paso a la Agricultura 4.0, donde tecnologías como el Internet de las Cosas (IoT), la inteligencia artificial (IA) y la automatización ciberfísica optimizan la producción y la sostenibilidad [1–4]. Sin embargo, esta transformación incrementa la vulnerabilidad a ciberataques, comprometiendo la disponibilidad, integridad y confidencialidad de los datos agrícolas, elementos esenciales para la toma de decisiones y la eficiencia operativa [3, 5, 6].

En este contexto, dos desafíos clave emergen: la ciberseguridad y la sostenibilidad. La creciente dependencia de infraestructuras digitales expone al sector a ataques dirigidos, filtraciones de datos y alteraciones en la información sensorial, con potenciales pérdidas económicas y de producción[5]. Paralelamente, la gestión eficiente del agua y la reducción del impacto ambiental se han convertido en prioridades, dado el impacto del cambio climático en la disponibilidad de recursos y la necesidad de prácticas agrícolas responsables [7–9].

La IA ha demostrado ser una herramienta clave en la automatización y predicción de variables agrícolas, permitiendo la detección temprana de enfermedades, la optimización del riego y la mejora en la aplicación de fertilizantes[7, 9–11]. Sin embargo, su eficacia depende de la protección robusta de datos y de infraestructuras de comunicación seguras, evitando manipulaciones malintencionadas que comprometan la toma de decisiones.

A pesar del avance en metodologías de seguridad, el sector agrícola aún adopta estándares genéricos como ISO 27000 y NIST CSF, sin una adaptación específica a sus limitaciones, como la baja conectividad en zonas rurales y la falta de formación en ciberseguridad del personal agrícola [5, 9, 12–14, 22]. Además, la fragmentación en la implementación de IA, sostenibilidad y ciberseguridad dificulta el despliegue de soluciones integrales [53].

Ante estos retos, presentamos QUILLAQUA, un marco diseñado para la gestión segura de cultivos hidropónicos. Su enfoque combina arquitecturas escalables de ciberseguridad, metodologías de análisis de riesgos y el uso eficiente de recursos mediante IA. QUILLAQUA busca ser una solución adaptable a distintos contextos agrícolas, garantizando productividad, resiliencia y sostenibilidad en un entorno digitalizado.

El artículo comienza con una visión general de los desafíos de la digitalización en la Agricultura 4.0, destacando la necesidad de integrar ciberseguridad, IA y sostenibilidad para garantizar la eficiencia y seguridad en la producción agrícola. En la Sección 2, se presenta una revisión del estado del arte, analizando estándares de ciberseguridad como ISO 27000 y NIST CSF, aplicaciones de IA en la optimización de cultivos y estrategias sostenibles enfocadas en la reducción del impacto ambiental. La Sección 3 desarrolla una revisión sistemática de la literatura, identificando vacíos en la investigación y consolidando hallazgos clave sobre la falta de un marco integral que combine estos tres pilares. En la Sección 4, se introduce y describe QUILLAQUA, un marco diseñado para la gestión segura y eficiente de cultivos hidropónicos, explicando su arquitectura, sus mecanismos de seguridad, el flujo operativo y su enfoque en la sostenibilidad. Finalmente, en la Sección 5, se presentan las conclusiones del estudio, resaltando la importancia de implementar soluciones escalables y adaptables en la Agricultura 4.0, además de delinear futuras líneas de validación y colaboración con entidades regulatorias para su aplicación en escenarios reales.

II. Estado de Arte

La Agricultura 4.0 se fundamenta en la incorporación de sensores, redes de comunicación, plataformas de Big Data

y sistemas avanzados de inteligencia artificial para la optimización de los procesos agrónomos. Estudios recientes han evidenciado que esta digitalización puede incrementar la eficiencia en el uso de insumos y mejorar la productividad, al tiempo que reduce el impacto medioambiental [7]. Sin embargo, el mismo proceso de conectividad y gestión de grandes volúmenes de datos introduce riesgos crecientes de ciberseguridad, lo que ha motivado la adaptación de estándares y metodologías procedentes de sectores industriales, como ISO 27000 o el marco NIST CSF, a los requerimientos del ámbito agrícola [15, 16].

En cuanto a ciberseguridad, en el contexto de las amenazas cibernéticas que ponen en riesgo la disponibilidad y fiabilidad de la información agraria —mediante ataques DoS o ransomware—, la necesidad de estrategias de resiliencia (planes de contingencia, segmentación de redes, políticas de parcheo y capacitación) se vuelve prioritaria. Sin embargo, el sector agrícola suele adoptar frameworks pensados para otros entornos industriales, requiriendo adaptaciones a sus condiciones (especialmente baja conectividad y dispersión geográfica) [17].

En la Tabla I, se resume estándares comunes, entre ellos ISO 27000 (orientado a SGSI), NIST CSF (modelo de identificación, protección, detección, respuesta y recuperación) y CIS Controls (guía practica de medidas básicas). Si bien cada uno ofrece pautas completas de seguridad, su implementación puede resultar costosa o compleja para operaciones rurales con recursos limitados. La clave está en adaptar dichos lineamientos a las particularidades agrarias: conectividad irregular, hardware robusto y personal no especializado. Solo así se garantiza la adopción sostenible de medidas de seguridad y la protección efectiva de la producción agrícola en el entorno de la Agricultura 4.0.

Tabla I
FRAMEWORKS FRECUENTEMENTE EMPLEADOS EN LA AGRICULTURA.

Framework	Características y Aplicabilidad
ISO 27000 Series [18]	**Relevancia:** Alta. Crítica para la protección de datos agrícolas sensibles. **Beneficios:** Certificación internacional, mejora continua, amplia cobertura de seguridad. **Limitaciones:** Complejo y costoso, especialmente para pequeñas operaciones agrícolas.
NIST Cybersecurity Framework (CSF) [6, 19, 20]	**Relevancia:** Muy alta. Ideal para infraestructuras críticas como la agricultura. **Beneficios:** Marco integral y adaptable que cubre desde la identificación de activos hasta la recuperación post-incidente. **Limitaciones:** Diseñado inicialmente para EE.UU., pero adaptable internacionalmente.
GDPR [21]	**Relevancia:** Alta para operaciones en la UE o que manejen datos de ciudadanos europeos. **Beneficios:** Mejora la confianza del consumidor y la protección de datos personales. **Limitaciones:** Requiere un enfoque riguroso en la protección de datos, lo que puede ser desafiante fuera de Europa.
NIST SP 800-53 [17, 22]	**Relevancia:** Amplio catálogo de controles de seguridad, adecuado para entornos altamente tecnológicos en agricultura digital.
CIS Controls [23]	**Relevancia:** Alta. Proporciona medidas de seguridad claras y prácticas. **Beneficios:** Fácil de entender y aplicar, enfoque práctico. **Limitaciones:** Menos reconocido internacionalmente que ISO o NIST.

La literatura más reciente propone, además, integrar la ciberseguridad como parte del diseño desde fases tempranas de los sistemas ciberfísicos, siguiendo un enfoque ”Security by Design” [24]. También está cobrando fuerza el término ciberbioseguridad, que abarca la protección de sistemas críticos de la cadena agroalimentaria y contempla la posibilidad de ataques dirigidos a alterar, de forma maliciosa, datos biológicos y genéticos [25]. En este sentido, se evidencia una carencia de marcos unificados que adopten un esquema específico para el entorno agrícola [57], frente a la práctica habitual de adaptar guías diseñadas para infraestructuras críticas de otro tipo.

Por otra parte, el empleo de inteligencia artificial (IA) ha mostrado resultados prometedores al proveer herramientas para la clasificación de semillas [14], la predicción de plagas y enfermedades [16] y el soporte en la toma de decisiones sobre riego y fertilización [15]. Mediante algoritmos de aprendizaje automático y técnicas de visión por computadora, es posible automatizar tareas tradicionalmente intensivas en mano de obra, además de anticipar condiciones ambientales adversas. No obstante, la IA demanda grandes volúmenes de datos, lo que conlleva nuevas vulnerabilidades si no se implementan controles adecuados para garantizar la confidencialidad, integridad y disponibilidad de la información [26, 27]. Asimismo, existe la necesidad de adaptar los modelos de IA a limitaciones reales de campo, como la variabilidad de los suelos o el acceso limitado a redes de banda ancha en zonas rurales.

En el ámbito de la sostenibilidad, la Agricultura 4.0 posibilita la adopción de prácticas de precisión, minimizando la huella hídrica y la aplicación excesiva de fertilizantes [28]. Sistemas como el riego inteligente, que ajusta automáticamente los volúmenes de agua en función de la humedad y la evapotranspiración, han mostrado ahorros considerables tanto en zonas áridas como en invernaderos de producción intensiva [24]. Igualmente, la implementación de métodos hidropónicos y aeropónicos permite cultivar con menor consumo de agua y menor contaminación de suelos, especialmente cuando se combina con el análisis predictivo de la IA [29]. Pese a ello, aún persiste la brecha de adaptar estas tecnologías de manera escalable y robusta, garantizando un balance entre rendimiento económico, seguridad alimentaria y cuidado medioambiental.

En síntesis, el estado del arte revela un escenario en rápida evolución, impulsado por la convergencia de tecnologías emergentes pero con importantes vacíos:

- Escasez de normas y marcos integrales que fusionen ciberseguridad, IA y sostenibilidad, diseñados de forma específica para la agricultura digital.
- Debilidad en la gestión de riesgos de ciberseguridad, especialmente en entornos con infraestructuras dispersas y personal poco habituado a protocolos de TI.
- Necesidad de metodologías formales que evalúen el impacto ambiental y validen las ganancias de productividad sin descuidar la seguridad de la información.
- Carencia de soluciones escalables que se ajusten a distintos contextos socioeconómicos, garantizando la aplicación de IA y medidas de protección independientemente de la disponibilidad de recursos tecnológicos.

Este panorama motiva la generación de enfoques multidisciplinares que aborden, de forma simultánea, la optimización productiva, la sostenibilidad ambiental y la seguridad de la información, pilares que inspiran la propuesta y el diseño del marco de trabajo que se presenta en este documento.

III. Revisión Sistemática

Esta sección expone los resultados de una revisión sistemática destinada a estudiar la convergencia de ciberseguridad, inteligencia artificial y sostenibilidad en Agricultura 4.0, con el fin de detectar vacíos, tendencias y soluciones frente a los desafíos de la modernización digital.

III-A. Protocolo y Marco de Referencia

Se adaptaron las pautas de Kitchenham y Brereton [30, 31] —usuales en ingeniería de software— al sector agrícola, definiendo:

- Definición de la pregunta de investigación:

  > RQ: "¿Qué investigaciones se centran en IA, ciberseguridad y sostenibilidad para impulsar la eficiencia y la resiliencia de los sistemas agrícolas ante retos contemporáneos?"

  La interrogante aborda protección de datos, uso de algoritmos y reducción del impacto ambiental.
- Diseño del protocolo: Se establecieron criterios de búsqueda, inclusión/exclusión y validación para estudios desde 2018.
- Estructura de la revisión: Estructura de la revisión: (1) Planificación (objetivos y palabras clave), (2) Ejecución (búsqueda bibliográfica y filtrado), (3) Análisis (calidad, extracción de datos y síntesis) y (4) Reporte (conclusiones y discusión).

III-B. Fuentes de Información y Estrategia de Búsqueda

Se consultaron bases de datos (IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Taylor & Francis, MDPI) y referencias manuales, usando conectores booleanos:

> ((Agriculture OR Smart Agriculture OR Urban agriculture OR Hydroponics OR Aeroponics OR Digital agriculture) AND NOT (Farm)) AND (Cybersecurity OR Artificial Intelligence) AND (Sustainability)

También se consideraron variantes (p. ej., "IoT Security", "digital farming").

III-C. Criterios de Inclusión y Exclusión

Se incluyeron estudios en inglés o español, revisados por pares y publicados desde 2018, que cubrieran al menos dos pilares (ciberseguridad, IA, sostenibilidad). Se excluyeron duplicados o trabajos sin aportes metodológicos ni componentes agroalimentarios o ambientales.

III-D. Proceso de Selección y Filtrado

De 370 resultados iniciales, quedaron 250 tras eliminar duplicados. Se revisaron títulos y resúmenes, quedando 80; tras su lectura, se seleccionaron 48 que cumplían los pilares y presentaban resultados tangibles. Se evaluó la calidad con una checklist, valorando coherencia, validez y pertinencia.

III-E. Extracción de Datos y Análisis

Se clasificaron los 48 artículos según categoría de contribución (Tabla II), extrayendo información bibliográfica, objetivos, amenazas de ciberseguridad, IA (algoritmos, métricas), sostenibilidad y limitaciones. Se aplicó minería de texto para evidenciar interacciones y vacíos.

Tabla II
Ciclo de Gestión Bibliográfica

Categoría	Cantidad	Referencia Bibliográfica
Metodología	4	[8, 9, 32, 33]
Framework	14	[4, 12, 20, 25, 26, 34, 34–40]
Modelo	9	[1, 16, 27, 41–44]
Técnica	11	[3, 5, 7, 10, 17, 19, 45, 46]
Proceso	2	[11, 47]
Otros	8	[2, 6, 15, 48–52]

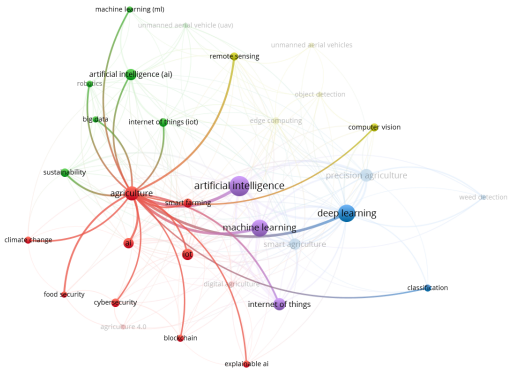


Figura 1. Mapa de conocimiento obtenido en agricultura

III-F. Síntesis de Hallazgos

En la Figura 1, "agriculture" se conecta con IA, IoT, machine learning y sustainability, reflejando la digitalización creciente. IA impulsa la automatización, mientras la ciberseguridad, asociada con blockchain e IoT, protege infraestructuras digitales. La sostenibilidad se vincula con climate change y food security, remarcando la necesidad de prácticas responsables.

Pese a ello, la adopción tecnológica es desigual (Figura 2). Solo 11 estudios contemplan ciberseguridad frente a 23 que la omiten. Treinta no consideran la gestión de la digitalización, y 43 no usan aprendizaje continuo, limitando la optimización de recursos. Solo 7 tienen estrategias sólidas de sostenibilidad, mientras 41 estudios ignoran soluciones de bajo coste. Además, 43 no incluyen casos prácticos y apenas 2 analizan la escalabilidad global.

Por ende, aunque la agricultura digital avanza en IA, IoT y análisis de datos, aún hay carencias en ciberseguridad, sostenibilidad y accesibilidad. Esto recalca la necesidad de marcos como QUILLAQUA, que integren protección de datos, eficiencia productiva y sostenibilidad para lograr una transformación digital más resiliente y escalable.

IV. Marco de Trabajo QUILLAQUA

La revisión sistemática evidenció la necesidad de un marco de trabajo que integre ciberseguridad, inteligencia artificial y sostenibilidad en la agricultura digital.

Este marco, denominado QUILLAQUA, se ha diseñado para abordar los desafíos de la Agricultura 4.0, asegurando un manejo inteligente de los recursos a través de la IA, una protección robusta de datos mediante estrategias de ciberseguridad, y un compromiso con la sostenibilidad ambiental. La imagen simboliza la unión de estos tres elementos, mostrando que el desarrollo tecnológico en la agricultura debe

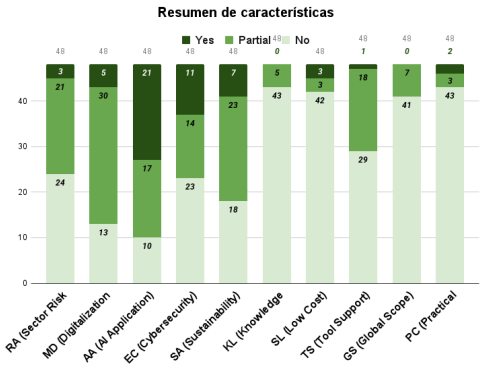


Figura 2. Resumen de características en la agricultura digital y sostenible.

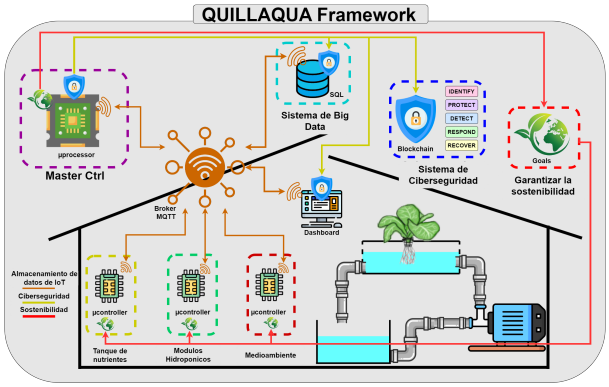


Figura 3. Descripción general de la propuesta.

ser equilibrado y resiliente, evitando riesgos en seguridad y maximizando el impacto positivo en el ecosistema.

En conclusión, la representación gráfica de QUILLAQUA refuerza la idea de un marco integral donde IA, sostenibilidad y ciberseguridad trabajan en conjunto, garantizando que la digitalización agrícola sea no solo eficiente y productiva, sino también segura y ecológicamente responsable.

A continuación, se ofrece una descripción detallada de QUILLAQUA, abarcando su arquitectura lógica, sus componentes fundamentales y las interacciones entre ellos, así como sus mecanismos de gestión de riesgos y su orientación a la sostenibilidad:

IV-A. Principios de Diseño

El marco QUILLAQUA ha sido diseñado con un enfoque integral que garantiza la seguridad, sostenibilidad y adaptabilidad en sistemas agrícolas digitalizados. Para ello, incorpora principios fundamentales que permiten su implementación en distintos entornos agrícolas, asegurando un equilibrio entre protección de datos, eficiencia operativa y optimización de recursos.

- Enfoque Integrado de Seguridad
  Siguiendo el principio de seguridad por diseño (Security by Design), QUILLAQUA incorpora estrategias de ciberseguridad desde la concepción del sistema agrícola. Esto incluye la identificación de activos críticos, evaluación de vulnerabilidades e implementación de controles adaptados al agro.
  El marco adapta estándares como el NIST Cybersecurity Framework (CSF) y la norma ISO 27000, ajustándolos a la agricultura digital, caracterizada por infraestructura distribuida, baja conectividad y manipulación por personal no especializado.
  La seguridad se aplica en todos los niveles del sistema, desde sensores básicos (pH, electroconductividad, humedad) hasta plataformas en la nube o servidores locales, garantizando protección de datos y resiliencia operativa ante amenazas físicas y digitales.
- Orientación a la Sustentabilidad
  QUILLAQUA prioriza la eficiencia en el uso de recursos y la reducción del impacto ambiental, asegurando una digitalización agrícola sostenible. Para ello:
  Integra IA para analizar consumo hídrico, huella de carbono y fertilización, ajustando dinámicamente las dosis según objetivos de sostenibilidad.
  Promueve la economía circular, reutilizando agua en sistemas hidropónicos con protocolos que evitan contaminaciones y protegen la calidad del cultivo.
  Optimiza el uso energético, fomentando sensores de bajo consumo y fuentes renovables en la infraestructura agrícola.
- IA Aplicada y Adaptativa
  QUILLAQUA incorpora aprendizaje continuo (online learning), que se ajusta dinámicamente a cambios en clima, variedad de cultivos y picos de consumo, optimizando la producción agrícola en tiempo real.
- Escalabilidad y Configurabilidad El marco es escalable tanto verticalmente (de pequeñas granjas a grandes fincas) como horizontalmente (incorporación de nuevos sensores y funciones). Además, ofrece configuración flexible, permitiendo ajustar políticas de seguridad y sostenibilidad según los recursos y capacidades de cada entorno.

IV-B. Arquitectura del Marco

El Framework QUILLAQUA representa la arquitectura del sistema diseñado para la gestión segura y sostenible de cultivos hidropónicos. La figura 3 se estructura en diferentes capas funcionales que interactúan entre sí para garantizar la seguridad de los datos, la eficiencia operativa y la sostenibilidad agrícola.

La Capa de Dispositivos y Sensores (Física), se incluyen sensores para medir pH, temperatura y humedad, junto con actuadores como bombas de riego y sistemas de iluminación controlados por dispositivos IoT que envían datos en tiempo real. Estos datos son gestionados por microcontroladores en los módulos de tanque de nutrientes, sistema hidropónico y entorno ambiental.

A nivel de ciberseguridad, se implementan cifrado de extremo a extremo, autenticación de dispositivos y actualizaciones de firmware, minimizando riesgos de manipulación o ataques a la infraestructura.

La Capa de Comunicaciones y Broker maneja la transmisión de datos entre la capa física y el núcleo de procesamiento mediante un Broker MQTT, asegurando cifrado y autenticación en la comunicación. Además, se incorporan mecanismos

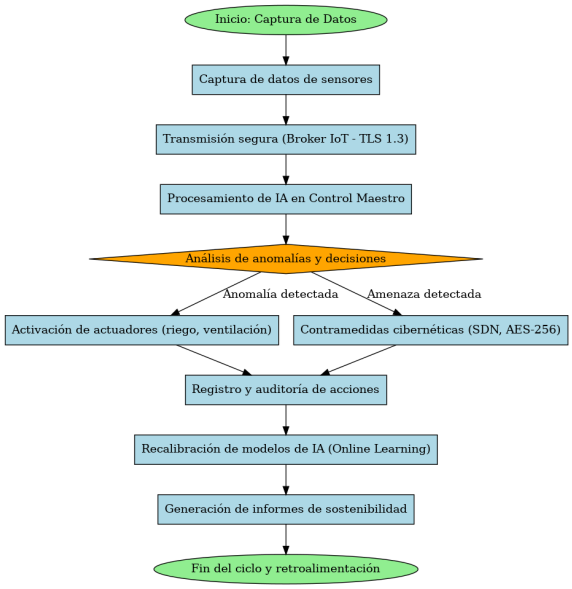


Figura 4. Flujo Operativo.

de monitoreo continuo para detectar tráfico anómalo o intentos de intrusión.

En la Capa de Procesamiento Principal, el Control Maestro (Master Ctrl), identificado en la imagen como un microprocesador, coordina la IA, administra la seguridad y supervisa la infraestructura. Los algoritmos de IA aplican técnicas como aprendizaje profundo y árboles de decisión, generando predicciones y sugerencias para optimizar cultivos.

Los Sistemas de Detección y Respuesta, representados en la imagen por el sistema de ciberseguridad basado en blockchain, protegen los datos mediante procesos de identificación, detección, protección, respuesta y recuperación ante amenazas digitales.

En la Capa de Aplicación/Base de Datos/Seguridad, la imagen muestra un Sistema de Big Data, donde los datos recopilados se almacenan y procesan para su visualización en dashboards de control. Esta capa gestiona accesos, cifrado de datos en reposo y auditorías de seguridad, garantizando el cumplimiento de políticas de protección.

Por último, en la Capa de Sostenibilidad, la imagen representa la sostenibilidad a través del módulo “Ensuring Sustainability”, que vincula el sistema hidropónico con el análisis de impacto ambiental. Esta capa mide huella hídrica, emisiones de CO_2 y eficiencia en fertilización.

El Módulo de Evaluación combina datos de IA y variables ambientales para optimizar el riego, reducir desperdicios y mejorar la eficiencia de los cultivos. También genera reportes periódicos para la toma de decisiones y el cumplimiento de normativas ambientales.

IV-C. Flujo Operativo

El ciclo operativo de QUILLAQUA, representado en la Figura 4, comienza con la captura dinámica de datos mediante sensores distribuidos en las unidades de cultivo. Estos sensores ajustan automáticamente sus períodos de muestreo en función de la etapa fenológica (germinación, floración) y los requerimientos del modelo de IA.

A continuación, los datos se transmiten de manera segura a través de un broker IoT que implementa cifrado TLS 1.3, autenticación mutua entre dispositivos y registros de metadatos (timestamp, geolocalización) para garantizar trazabilidad y prevenir alteraciones.

En la fase de procesamiento, el Control Maestro ejecuta algoritmos de IA (como LSTM para predicción hídrica o XGBoost para detección de anomalías en conductividad) en tiempo real, priorizando tareas críticas en entornos con conectividad intermitente.

Si se detectan desviaciones (pH anómalo, fluctuaciones térmicas), el sistema activa automáticamente los actuadores (bombas de riego, ventilación) y genera alertas en la interfaz de usuario. Paralelamente, se aplican contramedidas cibernéticas automatizadas como segmentación de red mediante SDN y rotación de claves AES-256, en respuesta a amenazas detectadas (inyección de datos falsos, ataques DDoS).

Cada acción ejecutada se registra en bases de datos temporales con marcas de auditoría, permitiendo una recalibración periódica de los modelos de IA mediante aprendizaje continuo (online learning), lo que mejora la precisión y adaptabilidad del sistema.

Además, el módulo de sostenibilidad, representado en la Figura 4, genera informes cuantitativos sobre métricas como litros de agua por kg producido y reducción de fertilizantes por ciclo, alineándose con estándares GRI e ISO 14064. Estos reportes proporcionan retroalimentación verificable, optimizando la seguridad, eficiencia operativa y responsabilidad ecológica.

Este flujo operativo, respaldado por la correlación de eventos con reglas MITRE ATT&CK adaptadas a la agricultura, garantiza una operación cohesiva donde cada capa del sistema refuerza a las demás, mitigando brechas identificadas en marcos convencionales de agricultura digital.

IV-D. Mecanismos de Gestión de Riesgos

En la fase inicial, se identifican activos y amenazas, clasificando dispositivos físicos (sensores, redes), digitales (bases de datos, software de IA) y lógicos (modelos, datos) según su criticidad. Luego, se realiza un análisis de vulnerabilidades, detectando fallos de configuración, firmware desactualizado o contraseñas débiles, y clasificando los riesgos según su impacto productivo, económico y ambiental (por ejemplo, una interrupción en el suministro de nutrientes puede generar pérdidas significativas).

Para la evaluación y tratamiento de riesgos, se adoptan estándares como NIST SP 800-53 e ISO 27005, priorizando amenazas y aplicando controles como cifrado, segmentación de red y autenticación de dos factores. QUILLAQUA documenta la relación amenaza-riesgo-control en una matriz, facilitando la trazabilidad de acciones correctivas.

Finalmente, el monitoreo continuo establece umbrales de seguridad y sostenibilidad (ej. niveles de fertilizantes), permitiendo que desviaciones activen planes de contingencia o ajustes en los modelos de IA. Este ciclo se repite durante toda la campaña agrícola, actualizando planes de respuesta a incidentes conforme evolucionan las amenazas y condiciones.

IV-E. Adaptabilidad y Escenarios de Aplicación

El marco QUILLAQUA está diseñado de forma modular para adaptarse a diversos entornos. En pequeñas explotaciones hidropónicas, donde los recursos de TI son limitados, ofrece una configuración simplificada con paneles intuitivos para monitoreo y detección de anomalías.

En fincas de agricultura de precisión a gran escala, permite la gestión de grandes volúmenes de datos con automatización avanzada, análisis predictivo y protocolos de seguridad (segmentación de redes por zonas de cultivo), asegurando escalabilidad sin comprometer el control de riesgos.

Para entornos urbanos (vertical farming o rooftop farming), optimiza iluminación LED y control climático mediante IA, incorporando indicadores de sostenibilidad como huella de carbono y consumo energético.

En zonas semiáridas o con escasez hídrica, se centra en la eficiencia del agua, integrando sensorización del suelo, riego optimizado y recirculación, con medidas de seguridad que evitan interrupciones o ataques que afecten la disponibilidad hídrica en momentos críticos.

IV-F. Sinergias entre Ciberseguridad, IA y Sostenibilidad

QUILLAQUA establece una interdependencia entre IA, ciberseguridad y sostenibilidad:

- IA y ciberseguridad: El análisis predictivo detecta anomalías en sensores o tráfico de red, adaptando dinámicamente las políticas de seguridad según la etapa del cultivo (ej. protección de datos en floración).
- Ciberseguridad y sostenibilidad: La protección de datos ambientales evita intrusiones que alteren registros de riego o fertilización, asegurando recomendaciones precisas de IA y evitando sobreconsumo de recursos.
- Sostenibilidad e IA: La IA equilibra rendimiento agronómico y huella ecológica, permitiendo configurar perfiles de sostenibilidad, como reducir un 15 % el uso de agua con solo 5 % de merma en producción, siempre dentro de umbrales de seguridad operativa.

Esta triada sinérgica, ausente en marcos convencionales, mitiga riesgos técnicos (ataques de false data injection) y garantiza que la productividad no comprometa la resiliencia ambiental, redefiniendo la agricultura digital como un ecosistema seguro y sostenible.

IV-G. Proyección, Validación del Marco y Aporte de QUILLAQUA

La validación de QUILLAQUA se plantea en dos fases:

1. Casos piloto en sistemas hidropónicos, midiendo respuesta ante ciberataques, reducción de incidentes, ahorro hídrico y uso óptimo de fertilizantes, con retroalimentación de agricultores y técnicos para optimizar interfaces y gestión de riesgos.
2. Pruebas a gran escala con cooperativas o grandes productores, evaluando escalabilidad, diversidad de parcelas y resiliencia ante ataques complejos.

Además, se busca alineación con estándares regulatorios (ISO 27000) y políticas de sostenibilidad para facilitar su adopción.

QUILLAQUA se distingue por integrar ciberseguridad proactiva, IA explicable y sostenibilidad cuantificable, con métricas como huella hídrica y emisiones de CO_2. Su diseño modular y adaptable permite aplicarlo en granjas hidropónicas o explotaciones extensivas, asegurando continuidad productiva, reducción ambiental y cumplimiento normativo. Esta sinergia tripartita lo posiciona como un marco pionero, abordando la brecha entre innovación tecnológica, seguridad localizada y responsabilidad ecológica en la agricultura 4.0.

V. Conclusiones

La convergencia tecnológica en la Agricultura 4.0 destaca la importancia crítica de la ciberseguridad y sostenibilidad para proteger datos y garantizar la producción. En primer lugar, estándares como NIST CSF e ISO 27000 son fundamentales para mitigar riesgos digitales, aunque requieren adaptaciones a las particularidades rurales: dispersión geográfica y limitaciones de conectividad.

En segundo término, la inteligencia artificial y análisis de datos avanzados optimizan recursos mediante modelos predictivos y visión por computadora, mejorando detección de plagas, reduciendo consumo hídrico y automatizando procesos. Sin embargo, el manejo de grandes volúmenes de datos exige reforzar la protección de información y gestión de identidades.

Como tercer pilar, la sostenibilidad demuestra su valor mediante agricultura de precisión y sistemas hidropónicos/aeropónicos, que aumentan productividad mientras reducen impacto ambiental. Esta sinergia se potencia cuando se integra con ciberseguridad, garantizando métricas confiables y procesos seguros.

El marco QUILLAQUA emerge como solución integral, combinando ciberseguridad, IA y sostenibilidad en arquitecturas modulares con enfoque Security by Design, adaptable desde pequeños cultivos hasta grandes explotaciones. Futuras acciones incluyen validación a escala, colaboración regulatoria y capacitación, avanzando hacia una agricultura digital segura y sustentable.

Agradecimientos

Este trabajo se ha desarrollado gracias a la financiación proporcionada por los siguientes proyectos: Proyecto Di4SPDS (CHIST-ERA grant- PCI2023145980-2) financiado por MCIN/AEI y cofinanciado por la Unión Europea, KOSMOS-UCLM (PID2024-155363OB-C44) financiado por MCIN/AEI/10.13039/501100011033; ALBA-UCLM (TED2021-130355B-C31) financiado por CIN/AEI/10.13039/501100011033/ Unión Europea NextGenerationEU/PRTR, y MESIAS (2022-GRIN-34202) financiado por FEDER.

Referencias

[1] G. Hu y F. You: “AI-enabled cyber-physical-biological systems for smart energy management and sustainable food production in a plant factory”, en *Applied Energy*, vol. 356, pp. 122334, 2024.

[2] A. K. Ng y R. Mahkeswaran: “Emerging and disruptive technologies for urban farming: A review and assessment”, en *Journal of Physics: Conference Series*, vol. 2003, n. 1, pp. 012008, 2021.

[3] S. P. Patil, L. M. Mathews, A. Kumar G., S. B. Motgi y U. Sinha: “AI-driven hydroponic systems for lemon

basil", en *International Conference on Network, Multimedia and Information Technology (NMITCON)*, pp. 1–6, 2023.

[4] R. R. A. Siregar, K. B. Seminar, S. Wahjuni y E. Santosa: "Vertical farming perspectives in support of precision agriculture using artificial intelligence: A review", en *Computers*, vol. 11, n. 9, 2022.

[5] J. Nikander, O. Manninen y M. Laajalahti: "Requirements for cybersecurity in agricultural communication networks", en *Computers and Electronics in Agriculture*, vol. 179, pp. 105776, 2020.

[6] J. Webb y D. Hume: "Campus IoT collaboration and governance using the NIST cybersecurity framework", en *Living in the Internet of Things: Cybersecurity of the IoT*, pp. 1–7, 2018.

[7] R. Abiri, N. Rizan, S. K. Balasundram, A. Bayat Shahbazi y H. Abdul-Hamid: "Application of digital technologies for ensuring agricultural productivity", en *Heliyon*, vol. 9, n. 12, pp. e22601, 2023.

[8] P. R, D. R y A. P: "Agriculture-based automation with recommendation systems based on AI models", en *Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 1582–1589, 2023.

[9] S. Sobue: "Space based data usage for smart farming", en *BIO Web of Conferences*, vol. 80, 2023.

[10] F. H. Juwono, W. K. Wong, S. Verma, N. Shekhawat, B. A. Lease y C. Apriono: "Machine learning for weed–plant discrimination in agriculture 5.0: An in-depth review", en *Artificial Intelligence in Agriculture*, vol. 10, pp. 13–25, 2023.

[11] J. Zscheischler, R. Brunsch, S. Rogga y R. W. Scholz: "Perceived risks and vulnerabilities of employing digitalization and digital data in agriculture – socially robust orientations from a transdisciplinary process", en *Journal of Cleaner Production*, vol. 358, pp. 132034, 2022.

[12] H. Chi, S. Welch, E. Vasserman y E. Kalaimannan: "A framework of cybersecurity approaches in precision agriculture", en *Proceedings of the International Conference on Cyber Warfare and Security*, p. 90, 2017.

[13] V. B. Sadu, N. Budidha, S. Billewar, S. J. Pimo, D. R. Kumar y T. C. A. Kumar: "Digitalization of agriculture by using the application of IoT and robotics", en *International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*, pp. 183–186, 2022.

[14] S. Sarkar, B. Ganapathysubramanian, A. Singh, F. Fotouhi, S. Kar, K. Nagasubramanian, G. Chowdhary, S. K. Das, G. Kantor, A. Krishnamurthy, N. Merchant y A. K. Singh: "Cyber-agricultural systems for crop breeding and sustainable production", en *Trends in Plant Science*, vol. 29, n. 2, pp. 130–149, 2024.

[15] U. Tariq, I. Ahmed, A. K. Bashir y K. Shaukat: "A critical cybersecurity analysis and future research directions for the internet of things: A comprehensive review", en *Sensors*, vol. 23, n. 8, 2023.

[16] S. K. V, D. B. Koppad, K. Awasthi, P. A. K y V. R.: "Application of AI models in agriculture", en *4th International Conference on Circuits, Control, Communication and Computing (I4C)*, pp. 387–390, 2022.

[17] N. Munga y N. Naidoo: "Cyber Security in Smart Agriculture Technology", en *Snode Technologies*, 2023.

[18] ISO-IEC-27000: "Information Technology Security Techniques - Information Security Management Systems - Overview and Vocabulary", International Organization for Standardization, 2018.

[19] Z. Angyalos, S. Botos y R. Szilagyi: "The importance of cybersecurity in modern agriculture", en *Journal of Agricultural Informatics*, vol. 12, n. 2, 2022.

[20] J. West: "A prediction model framework for cyber-attacks to precision agriculture technologies", en *Journal of Agricultural & Food Information*, vol. 19, n. 4, pp. 307–330, 2018.

[21] General-Data-Protection-Regulation-GDPR: "General data protection regulation (GDPR)", 2016.

[22] W. L. Ross Jr.: "Security and Privacy Controls for Information Systems and Organizations", NIST, 2020.

[23] CIS-Critical-Security-Controls: "CIS Critical Security Controls® Version 8.1", 2021.

[24] M. Hobbins, O. Boiko, C. Dewes, A. Hoell, G. Husak, H. Jayanthi, T. Magadzire, A. McNally, D. Sarmiento, G. Senay, et al.: "A new global reference evapotranspiration reanalysis: global opportunities in operational drought monitoring and famine early warning", en *EGU General Assembly Conference Abstracts*, pp. EGU–16989, 2023.

[25] S. E. Duncan, R. Reinhard, R. C. Williams, F. Ramsey, W. Thomason, K. Lee, N. Dudek, S. Mostaghimi, E. Colbert y R. Murch: "Cyberbiosecurity: A new perspective on protecting U.S. food and agricultural system", en *Frontiers in Bioengineering and Biotechnology*, vol. 7, 2019.

[26] A. A. Aliyu y J. Liu: "Blockchain-based smart farm security framework for the Internet of Things", en *Sensors*, vol. 23, n. 18, pp. 7992, 2023.

[27] M. S. Dimitrijević: "Technological progress in the function of productivity and sustainability of agriculture: The case of innovative countries and the republic of serbia", en *Journal of Agriculture and Food Research*, vol. 14, pp. 100856, 2023.

[28] M. Baatsen, P. Bijl, A. von der Heydt, A. Sluijs y H. Dijkstra: "Resilient antarctic monsoonal climate prevented ice growth during the eocene", en *Climate of the Past*, vol. 20, n. 1, pp. 77–90, 2024.

[29] C. Colesie, C. V. Walshaw, L. G. Sancho, M. P. Davey y A. N. Gray: "Antarctica's vegetation in a changing climate", en *WIREs Climate Change*, vol. 14, 2022.

[30] B. Kitchenham y P. Brereton: "A systematic review of systematic review process research in software engineering", en *Information and Software Technology*, vol. 55, n. 12, pp. 2049–2075, 2013.10.1016/j.infsof.2013.07.010.

[31] S. Barat, T. Clark, B. Barn and V Kulkarni: "A model-based approach to systematic review of research literature", en *Proceedings of the 10th Innovations in Software Engineering Conference*, 2017.

[32] R. Ramyashree y V. Rao A. S: "Application of AI technology and image processing to smart agriculture: Detection and classification of plant diseases", en *13th International Conference on Computing Communication*

and Networking Technologies (ICCCNT), pp. 1–6, 2022.

[33] G. Singh, N. Kalra, N. Yadav, A. Sharma y M. Saini: "SMART AGRICULTURE: A REVIEW", en *Siberian Journal of Life Sciences and Agriculture*, vol. 14, n. 6, pp. 423–454, 2022.

[34] S. Alyahya, W. U. Khan, S. Ahmed, S. N. K. Marwat y S. Habib: "Cyber secure framework for smart agriculture: Robust and tamper-resistant authentication scheme for IoT devices", en *Electronics*, vol. 11, n. 6, pp. 963, 2022.

[35] P. Blanco-Carmona, L. Baeza-Moreno, E. Hidalgo-Fort, R. Martín-Clemente, R. González-Carvajal y F. Muñoz-Chavero: "AIoT in agriculture: Safeguarding crops from pest and disease threats", en *Sensors*, vol. 23, n. 24, 2023.

[36] H. T. Bui, H. Aboutorab, A. Mahboubi, Y. Gao, N. H. Sultan, A. Chauhan, M. Z. Parvez, M. Bewong, R. Islam, Z. Islam, S. A. Camtepe, P. Gauravaram, D. Singh, M. A. Babar y S. Yan: "Agriculture 4.0 and beyond: Evaluating cyber threat intelligence sources and techniques in smart farming ecosystems", en *Computers and Security*, vol. 140, pp. 103754, 2024.

[37] S. S. Panda, S. Banerjee y S. Alok: "Application of industry 4.0 technologies in climate-smart agricultural practices", en *Fostering Sustainable Development in the Age of Technologies*, pp. 289–302, 2023.

[38] P. Savvidis y G. A. Papakostas: "Remote crop sensing with IoT and AI on the edge", en *IEEE World AI IoT Congress (AIIoT)*, pp. 0048–0054, 2021.

[39] C. Vuppalapati, A. Ilapakurti, S. Kedari, J. Vuppalapati, S. Kedari y R. Vuppalapati: "Democratization of AI, albeit constrained IoT devices y Tiny ML, for creating a sustainable food future", en *3rd International Conference on Information and Computer Technologies (ICICT)*, pp. 525–530, 2020.

[40] Á. J. Varela-Vaca, D. G. Rosado, L. E. Sánchez, M. T. Gómez-López, R. M. Gasca y E. Fernández-Medina: "CARMEN: A framework for the verification and diagnosis of the pecification of security requirements in cyber-physical systems", en *Computers in Industry*, vol. 132, pp. 103524, 2021.

[41] M. Kabir y S. Ekici: "Energy-agriculture nexus: Exploring the future of artificial ntelligence applications", en *Energy Nexus*, vol. 13, pp. 100263, 2024.

[42] B. Kusumasri, S. V, S. Satyavada y G. U. Kiran: "Crop recommendation application using ensemble classifiers", en *2nd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, pp. 1–7, 2023.

[43] A. Rehman, I. Abunadi, K. Haseeb, T. Saba y J. Lloret: "Intelligent and trusted metaheuristic optimization model for reliable agricultural network", en *Computer Standards & Interfaces*, vol. 87, pp. 103768, 2024.

[44] S. Silva, D. Duarte, A. Valente, S. Soares, J. Soares y F. C. Pinto: "Augmented intelligent distributed sensing system model for precision agriculture", en *Telecoms Conference (ConfTELE)*, pp. 1–4, 2021.

[45] R. Das, S. S. Bhatt, S. Kathuria, R. Singh, G. Chhabra y P. K. Malik: "Artificial intelligence and internet of things based technological advancement in domain of horticulture 4.0", en *IEEE Devices for Integrated Circuit (DevIC)*, pp. 207–211, 2023.

[46] M. Nery, R. Santos, W. Santos, V. Lourenco y M. Moreno: "Facing digital agriculture challenges with knowledge engineering", en *First International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 118–119, 2018.

[47] T. Drape, N. Magerkorth, A. Sen, J. Simpson, M. Seibel, R. S. Murch y S. E. Duncan: "Assessing the role of cyberbiosecurity in agriculture: A case study", en *Frontiers in Bioengineering and Biotechnology*, vol. 9, 2021.

[48] N. Al Hashlamoun, N. Al Barghuthi y H. Tamimi: "Exploring the intersection of AI and sustainable computing: pportunities, challenges, and a framework for responsible applications", en *9th International Conference on Information Technology Trends (ITT)*, pp. 220–225, 2023.

[49] M. A. Rather, I. Ahmad, A. Shah, Y. A. Hajam, A. Amin, S. Khursheed, I. Ahmad y S. Rasool: "Exploring opportunities of artificial intelligence in aquaculture to meet increasing food demand", en *Food Chemistry: X*, vol. 22, pp. 101309, 2024.

[50] Pallavi B, B. Othman, G. Trivedi, N. Manan, R. S. Pawar y D. P. Singh: "The application of the internet of things (IoT) to establish a technologically advanced industry 4.0 for long-term growth and development", en *2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 1927–1932, 2022.

[51] S. Souravlas, N. Tantalaki y M. Roumeliotis: "Data-driven decision making in precision agriculture: The rise of big data in agricultural systems", en *Journal of Agricultural & Food Information*, vol. 20, n. 4, pp. 344–380, 2019.

[52] K. K. Ramachandran, B. Nagarjuna, S. V. Akram, J. Bhalani, A. M. Raju y R. Ponnusamy: "Innovative cyber security solutions built on blockchain technology for industrial 5.0 applications", en *International Conference on Artificial Intelligence and Smart Communication (AISC)*, pp. 643–650, 2023.

[53] G. Büyüközkan y D. Uztürk: "Integrated design framework for smart agriculture: Bridging the gap between digitalization and sustainability", en *Journal of Cleaner Production*, vol. 449, pp. 141572, 2024.

[54] FAO Regional Office for Europe and Central Asia: "Global agriculture towards 2050. In High Level Expert Forum - How to Feed the World in 2050", 2023.

[55] F. Martínez, L. E. Sánchez, A. Santos-Olmo, D. G. Rosado y E. Fernández-Medina: "Maritime cybersecurity: Protecting digital seas", en *International Journal of Information Security*, pp. 1–29, 2024.

[56] S. Qazi, B. A. Khawaja y Q. U. Farooq: "IoT-equipped and AI-enabled next generation smart agriculture: A critical review, current challenges and future trends", en *IEEE Access*, vol. 10, pp. 21219–21235, 2022.

[57] M. Campoverde-Molina y S. Luján-Mora: "Cybersecurity in smart agriculture: A systematic literature review", en *Computers & Security*, p. 104284, 2024.

Sesión Investigación IV: Seguridad en sectores estratégicos

Assessing cybersecurity of Internet-facing medical IT systems in Germany & Spain using OSINT tools

Pere Tuset-Peiró
TecnoCampus - Universitat Pompeu Fabra
Av. Ernest Lluch, 32
08302 Mataró, Spain
ptuset@tecnocampus.cat

Michael Pilgermann
Brandenburg University of Applied Sciences
Magdeburger Str. 50
14770 Brandenburg an der Havel, Germany
michael.pilgermann@th-brandenburg.de

Josep Pegueroles
Universitat Politècnica de Catalunya
C/Jordi Girona, 1-3
08034 Barcelona, Spain
josep.pegueroles@upc.edu

Xavier Vilajosana
Universitat Oberta de Catalunya
Rambla del Poblenou 156
08017 Barcelona, Spain
xvilajosana@uoc.edu

***Abstract*—This paper investigates cybersecurity threats in medical IT (Information Technology) systems exposed to the Internet. To that end, we develop a methodology and build a data processing pipeline that allows to gather data from different OSINT (Open Source Intelligence) sources, and processes it to obtain relevant cybersecurity metrics. To validate its operation and usefulness, we apply it to two countries, Germany and Spain, allowing to study the main threats that affect medical IT systems in these countries. Our initial findings reveal that 20% of German hosts and 15% of Spanish hosts tagged as medical devices have at least one CVE (Common Vulnerabilities and Exposures) with a CVSS (Common Vulnerability Scoring System) graded as critical (i.e., value 8 or greater). Moreover, we find that 74% of CVEs found in German hosts are dated from earlier than 2020, whereas for Spanish hosts the percentage is 60%. This indicates that medical IT systems exposed to the Internet are seldom updated, which further increases their exposure to cyberthreats. Based on these initial findings, we finish the paper providing some insights on how to improve cybersecurity of these systems.**

***Index Terms*—Medical IT systems, PACS, DICOM, OSINT.**

Tipo de contribución: *Investigación en desarrollo*

I. Introduction

In the early 1980s, medical equipment manufacturers pioneered interoperability by developing standards, such as ACR/NEMA 300, to facilitate data exchange between CT (Computed Tomography) and MRI (Magnetic Resonance Imaging) systems. The advent of the Internet in the early 1990s further enhanced communication and management of medical devices, with medical image protocols adopting the TCP/IP stack. At that time DICOM (Digital Imaging and Communications in Medicine) emerged as a prominent protocol enabling storage, retrieval, and printing of medical images across various devices [1]. To further ease the integration of DICOM devices with the Internet, PACS (Picture Archiving and Communication Systems) systems were developed in the late 1990s , which integrate a DICOM interface to retrieve data from medical devices and provide a common HTTP (Hyper-Text Transfer Protocol) interface to expose data to medical staff and patients through a web browser [2].

Such approach has clearly increased the efficiency in patient care, but it has also introduced significant cybersecurity risks, specially when considering that integrating security mechanisms to DICOM and HTTP was an afterthought. This has set the stage for evolving cybersecurity risks in the medical domain, specially when considering that these medical IT systems are now being exposed to the Internet. For example, in 2016 researchers discovered 2774 DICOM servers exposed to the Internet worldwide [3]. In 2019 researchers from Greenbone Networks [4] showed that many PACS servers worldwide are accessible via the Internet without any security measures. In 2023 Sina Yazdanmehr from aplite GmbH [5] also showed in a BlackHat conference that DICOM servers and HTTP servers used in medical IT systems are largely exposed to the Internet, which can lead to patient data leaks. Finally, in 2024 Censys [6] identified over 14000 unique IP addresses exposing medical-related services, including 5100 DICOM and 4000 EMR/EHR (Electronic Medical Records / Electronic Health Records) services. Hence, it is urgent to comprehensively assess the cybersecurity status of Internet-exposed medical IT systems world-wide.

Given these preliminary findings, this study aims to propose a flexible and open data acquisition and processing pipeline that enables the collection and analysis of information from healthcare systems connected to the Internet. The proposed methodology will allow for the investigation of vulnerabilities in medical IT systems connected to the Internet based on OSINT (Open Source Intelligence) tools. We decide to use OSINT tools, such as Censys and Shodan, since its usefulness has already been demonstrated to study vulnerabilities of Internet-facing systems, such as power grid systems [7]. Using the developed pipeline, we focus this initial study on German and Spanish medical IT systems, allowing to validate its correct operation and to gather evidence about their status. However, the methodology and the tools employed in this study can serve as a foundation for wider studies assessing global medical IT cybersecurity, which we plan to do in the future. Last but not least, it is important to disclose that during the development of this project no active hacking has been conducted. Moreover, all the results that have been found are in the process of being reported to the stakeholders, process that will be completed before the publication of this study.

The remaining of the paper is organized as follows. Sec-

tion II presents an overview of cybersecurity threats faced by medical IT systems. Section III details the methodology that we have followed and the tools that we have developed to conduct this study. Section IV presents and discusses the results that have been obtained as part of the study. Finally, Section V summarizes our findings and outlines future work.

II. Cybersecurity in the medical domain

There is limited generic material available on how hospitals look with respect to their business processes and their technological implementation.

In general, (networked) systems in a hospital are grouped along the following domains: IT (Information Technology), MT (Medical Technology), and utility systems. In turn, the systems in the IT domain can further be divided into two categories: general administrative IT (PCs and servers that support administrative tasks and communication) and specialized hospital IT systems, including HIS (Hospital Information System), RIS (Radiology Information System), LIS (Laboratory Information System), and PACS (Picture Archiving and Communication Systems). MT in hospitals includes a wide range of critical devices used for diagnosing, monitoring, and treating patients. These include ventilators, infusion pumps, patient monitoring and imaging systems such as MRI (Magnetic Resonance Imaging) and CT (Computed Tomography) scanners. Due to their growing network connectivity, these devices are increasingly referred to as Medical Cyber-Physical Systems (MCPS) in the literature [8].

The heterogeneity of systems across the IT and MT domains in hospitals creates significant complexity. Furthermore, the interconnection of these devices within and between these domains (including the Internet) offers numerous operational advantages, but also introduces new security vulnerabilities that must be carefully managed to ensure the safety of both patients and staff. Hence, detecting and patching vulnerabilities throughout the medical equipment in a hospital has become a top priority for the staff, but the teams managing such processes are often non-existing or overloaded, thus delaying such procedures.

1) Hospitals IT cybersecurity status: Sources like Check Point Research publications [9] report that the healthcare sector is now among the three most attacked ones, and ENISA (the European Union Agency for Cybersecurity) claims that hospitals are particularly affected in terms of cybersecurity, as RansomWare remains the predominant threat, driven by the prospect of financial gain [10]. Following reports of the Health Sector Cybersecurity Coordination Center, the interest of attackers has increased to the point that the that the top ten ransomware groups actively target the US health and public health sector [11].

2) Vulnerabilities and their management: Regarding to [12], 33.524 vulnerabilities were recorded in the NIST NVD between July 2023 and June 2024, which makes up for an increase of 35 % compared to the reporting period one year before (24.690 vulnerabilities). For organisations it is challenging to identify the most relevant vulnerabilities and prioritize countermeasures. To address these issues, VM (Vulnerability Management) was developed [13].

III. Methodology

In this section, we present the proposed methodology, along with the existing tools selected and those developed specifically for this work. In each corresponding subsection the detailed steps of each phase is described in detail.

A. Scanning the Internet for medical IT systems

Scanning the Internet to discover services running on hosts has been one of the main ways to uncover vulnerabilities in different environments, albeit a slow one. Indeed, scanning the whole range of IPv4 addresses is an endeavour that can take various weeks if done sequentially given the large number of hosts (i.e., $2^{32} = 4294967296$), ports (i.e., $2^{16} = 65536$) and protocols (i.e., TCP and UDP) to explore. The process can be speed up by splitting the work among various servers and running it in parallel, but it is still a slow process given the magnitude of the search space. In addition, while scanning hosts to detect services is not considered a crime in the United States and Europe, it can still cause legal problems when such procedure is not explicitly authorized by the administrators. Moreover, active scanning is considered an attack by many SOCs and cause alarms when done massively.

Since we decided to adhere to a strict no-active hacking for the development of the project, we use a methodology based on OSINT which, according to its pioneer Robert David Steel, is "*the art and science of creating ethical, evidence-based decision support using only open sources and methods—legal and ethical in every respect*" [14]. Hence, we decided to use search engines that periodically scan the whole range of IPv4 addresses and store the services running in each host, as well as related metadata (i.e., geographical information, DNS information, etc.), allowing to speed up the process and save computing and network resources. Based on the information already stored in their database, these services allow users to perform queries and filter the results to search for specific information. For example, it is possible to query the database to discover all the ports open on a specific IP addresses, or to query for any IP address that is running a certain service. Moreover, since the information stored in the database is augmented with context information (i.e., IP geo-location), it is also possible to query by the country where the hosts are running, among others.

The OSINT tools that we aim to integrate in our processing pipeline should meet the following requirements: 1) Well-established within the community, providing reliable results, 2) Provide basic information (i.e., ports, services, etc.) and additional metadata (CVEs, geoposition, etc.), 3) Provide frequent updates about the hosts and the services they are running, and, 4) Provide an API to automate the process of making queries and retrieving results.

As of today, there are two main OSINT tools that meet the described criteria: Shodan (https://www.shodan.io/) and Censys (https://censys.com/). The former was founded by John Matherly in 2009, whereas the latter was founded by Zakir Durumeric in 2017, one of the developers of ZMap [15], a collection of open source tools to perform large-scale studies of Internet hosts and services. Shodan provides more detailed information for each host and service, like a list of CVEs, which is very useful to report found vulnerabilities. In contrast, Censys has a better query and filter system, allowing to detect hosts and services with more precision.

B. Evaluating OSINT tools through a hackathon

In order to explore the problem at hand and evaluate the OSINT tools, we organized a hackathon as an activity of the

CARISMATICA project, led by UPC (Universitat Politècnica de Catalunya) and funded by INCIBE (Instituto Nacional de Ciberseguridad). The hackathon took place at TecnoCampus Mataró in December 2024, with the participation of 18 students from various universities in Germany and Spain. The participants had different backgrounds, including Computer Science and Electrical Engineering, and different academic levels, from Bachelor's to Master's degree.

When preparing the hackathon we devised two approaches to kick off the data collection and processing pipeline:

- *A-priori methodology*: The process starts with a preparation phase, in which IP ranges linked to medical organizations are identified using additional OSINT sources (i.e., an atlas of medical institutions in each country), so that the follow-up steps of the pipeline only work on a subset of the complete public IP address range.
- *Zero-knowledge methodology*: The process starts with a Censys request and solely relies on the tagging mechanisms provided by Censys to find IP addresses linked to medical IT devices. Afterwards, the information is enriched with other OSINT tools, such as Shodan, to obtained a detailed overview of each IP address.

During the hackathon, students were mixed in 4 groups and worked in 2-hours sprints over the course of 4 days, with the only rule that no active hacking was allowed. In the first sprint, we provided some general ideas on the objective and the methodology, as well as access to the OSINT tools described earlier, and allowed them to explore the methodologies and the APIs of each tool. After the first sprint each group selected a methodology, which they would use for the following sprints, and started developing code to implement the data collection and processing pipeline. At the end of each sprint each group of students provided a short update of their progress and any relevant findings, and at the end of the hackathon each group presented their results.

Overall, the results of the hackathon were very positive, both in terms of student engagement and practical outcomes. First and foremost, we could witness how students started from a very basic idea of the goal and almost no knowledge of the tools, to being able to develop code that allowed to get some initial results. It is needless to say that the moment that students got their first results, the energy and the motivation in the room boomed. Second, it allowed to validate our ideas on the data collection and processing pipeline, and to validate the potential of the project. Hence, we believe that hosting a hackathon was a very good experience for everyone involved, and we plan to repeat it next year.

C. Developing a data acquisition and processing pipeline

As we learned through the hackathon, collecting and processing data from OSINT tools involves a huge amount of data from different sources. To simplify data handling and make results reproducible, we have developed a set of processing pipelines using the Python programming language. Several processing steps have to be combined, usually by piping information between several Python scripts or persist output from a processing step into SQLite databases, CSV files, or JSON files. The processing pipelines have in common that they are meant to work on a set of IP addresses, which are linked to medical IT devices, and run (semi-automatically) several OSINT procedures, allowing to identify and report on relevant security information for each host.

Based on the experience acquired through the hackathon, we decided to use the *zero knowledge* methodology. Hence, we start by using Censys to collect the IP addresses of hosts in each country that are tagged as '`medical-device`'. To that end, we developed a Python script that uses the Censys API search feature to run the query '`location.country: 'country' and labels='medical-device''`, where `country` is a variable indicating the country (i.e., Germany or Spain). The results returned by Censys are in the form of a JSON structure with the IP addresses of devices that meet the search criteria. After that, we use the Censys API view feature to retrieve all the information for each host. For each host, the Censys API returns a JSON structure that contains information about the host in the form of a dictionary. The keys of the dictionary point to other data structures that can be other dictionaries or lists. At the first level of the structure, the dictionary includes keys such as `ip`, `services`, `location`, `whois` and `dns`, among others. Then, within the `services` key we find a list of dictionaries, each containing further information about each services, such as, the name of the service (`service_name`) and the port in which the service is running (`port`), among others. All the information of each host retrieved in each query is stored in a file that is named after the IP address of the host and with the `.censys` extension. The files are stored under a directory named after the country, i.e., '`Spain/IP_address.censys`'.

Once all the hosts for each country have been retrieved using Censys, another Python script uses the Shodan API host feature to query information about that host. The information provided by Shodan is structured in a similar way (i.e., dictionary containing dictionaries or lists), but provides different information. For example, the Shodan query returns a first level dictionary with a summary of the collected information (i.e., IP address, ISP, geolocation, ASN, etc.) and another dictionary with all the raw data. Again, the information retrieved from Shodan is stored in a file named after the IP and with the `.shodan` extension under the country directory, i.e., '`Spain/IP_address.shodan`'. During the process we have found that in some cases Shodan does not have information for a given host provided by Censys, in which case the files are not created. This is accounted for when further processing the data files of each country to calculate statistics and plot the results.

Once all the data for each country has been retrieved from Censys and Shodan, another Python script is responsible to analyse the services and ports running in each host and to grab screenshots of the HTTP services using the Selenium library. The screenshots of each host are stored in a PNG file, allowing to manually analyse the obtained results for each host/port and determine which medical IT services are running on each.

Finally, another set of Python scripts in the data processing pipeline are responsible to analyse the data, compute statistics and create the plots, allowing us to analyse the potential impact of vulnerabilities, compare the state of cybersecurity in each country, and to draw conclusions. In the process of analysing the potential impact of vulnerabilities, it is important to mention that we use the CVE (Common Vulnerabilities and Exposure) system and the CVSSv3 (Common Vulnerability Scoring System, version 3) score. On the one hand, the CVE is a unique number that uniquely identifies

each discovered vulnerability, either in software or hardware. On the other hand, the CVSS is a number between 0 and 10 that is assigned to each CVE depending on its severity level, ranging from low (0.0-3.9) to critical (9.0 to 10.0). To find information about the CVE and the CVSS we use the NVD (National Vulnerability Database), which receives its information directly from MITRE and is considered to be a primary source [16], [17].

IV. Results

This section presents and analyses the results that have been obtained using the methodology described earlier in the paper.

A. Dataset overview

The procedure to gather information from German and Spanish hosts tagged as medical devices was conducted the third week of February 2025. It is important to mention that the process of gathering the data is slow, since both Censys and Shodan impose a limit of around one query per second and the lists of host can be large (i.e., over 1000 per country).

Using the methodology described earlier, Censys returned a total of 54 hosts for Spain, and a total of 1449 hosts for Germany. In turn, Shodan returned 49 hosts out of the 54 hosts (90.7%) for Spain, whereas it returned 1426 hosts of the 1449 hosts (98.4%) for Germany.

Based on the obtained results, we start our dataset overview by checking on information freshness of each service by comparing the `last_updated_at` tag on Censys and the `last_update` tag on Shodan. For Spain the results show that 75% of the time information provided by Censys is more up-to-date than Shodan. In contrast, for Germany the results show that 75% of the time Shodan provides more up-to-date information than Censys.

We also check the size of the information (i.e., JSON structure) returned by each service for each host. The results for Spain show that on average Censys returns two times (2x) more information for each host than Shodan, whereas for Germany Censys returns six times (6x) more information for each host than Shodan. However, for some hosts we have found that Shodan provides more detailed information, such as a list of the CVEs affecting the hosts.

Finally, we also show were hosts are located in each country, as depicted in Figure 1. The black dots in the maps show the (approximate) location of each hosts, whereas the heat-map shows the percentage of hosts found in each region of the country with respect to the total number of hosts detected in the country. As it can be observed, the regions of Hessen in Germany and Madrid in Spain are the ones with a higher percentage of vulnerable hosts found, with a percentage of 90% and 70%, respectively. Such concentration can be explained by the fact that most services are hosted by cloud providers, such as Amazon (i.e., the `3.64.0.0/12` IP range is announced from Hessen), or Internet Service Providers, such as Telefonica (i.e., in Madrid).

B. Dataset cleaning

To analyse the potential cybersecurity impact on medical devices in Germany and Spain, we start by focusing on the ports that are open for every host. Analysing the dataset we notice that in Germany the average number of open ports per host is 52.86 ports, with a standard deviation of 56.88, and the host with the most number of open ports is 582. In contrast, in Spain the average number of open ports per host is 7.94, with a standard deviation of 7.35, and the host with the most number of open ports is 37.

Having hosts with such large number of open ports makes us suspicious that a fraction of hosts may be honeypots. Hence, we start by plotting the percentage of hosts with a given number of open ports for Germany (red) and Spain (blue). The results depicted in Figure 2 confirm our suspicion. As it can be observed, in Germany the distribution shows two clear groups of hosts: a group with a number of open ports between 1 and 40, and another group with more than 40 open ports. In contrast, in Spain all hosts have less than 30 open ports except one. However, only relying on the number of open ports may lead to some hosts being flagged as false positives (i.e., regular hosts flagged as honeypots) or false negatives (i.e., honeypots flagged as regular hosts).

To further confirm our hypothesis, we analyse the hosts in Germany to determine where the IPs of the hosts are being originated from. As depicted in Figure 3, we observe that most hosts that have more than 30 open ports are originated from ASN 16509, which belongs to AWS (Amazon Web Services).

In addition to that, we also check the values returned from the reverse DNS query on the IP address. Using such approach, hosts that are not honeypots should only have a handful of DNS records to allow humans access the services that are hosted. In contrast, hosts that are honeypots should have none or only one DNS record, since humans are not expected to access the services that are hosted anyway. Moreover, as we have seen earlier, hosts that are honeypots will most probably be announced from a cloud service provider, such as AWS. As depicted in Figure 4, hosts in Germany with over forty open ports and zero or one DNS record are mostly announced from AWS, confirming our hypothesis.

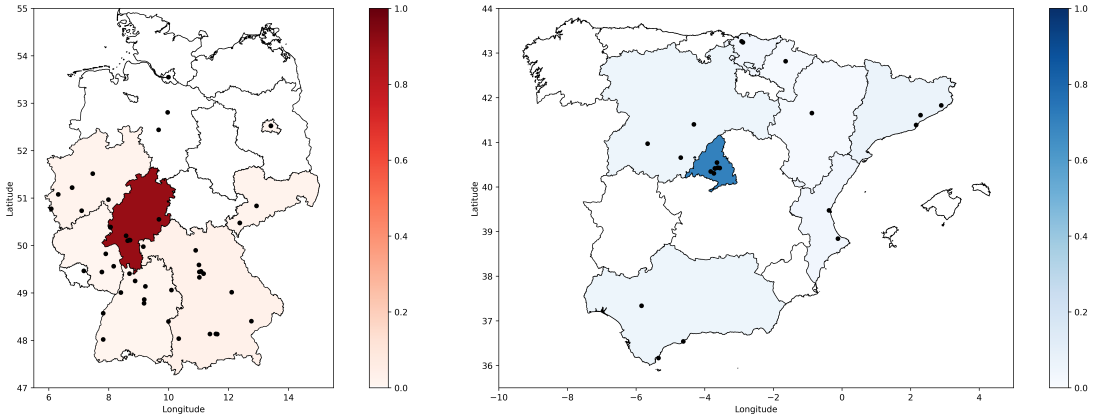
Hence, seeing the strong correlation between the three approaches, we have decided to remove hosts with more than 30 open ports from the remaining of our analysis, since they can have a huge impact on the results and the conclusions.

C. Dataset analysis

After removing the hosts flagged as potential honeypots, we now focus on the services, as determined by the protocol that is executing (and not the port where the socket is listening), that are run by hosts in Germany (red) and Spain (blue), as depicted in Figure 5. As it can be observed the most used protocol in hosts tagged as `medical-devices` by Censys is HTTP, with 53% for Germany and 42% for Spain. The remaining results for Germany and Spain are similar, with hosts running SSH, DICOM, FRP, IMAP, MYSQL, POP3 and SMTP, among others. In addition, there are also a 12.8% of ports that are classified as unknown.

Having HTTP services exposed to the Internet is normal as long as such services are intended for end-users accessing it, and as long as the services are implemented following appropriate coding guidelines. What is not normal is having so many services running on such hosts. For example, we observe some hosts exposing database or email services.

Next, we focus on analysing the operating system used by hosts, since this is another potential threat surface, specially for hosts running older operating systems that may not be patched. As displayed in Figure 6, in Germany we observe that 35% of the hosts use Linux, whereas Windows only represents 4% of the hosts. In contrast, in Spain Linux represents 28%



(a) Germany map

(b) Spain map

Figure 1: Heat-map of Germany and Spain showing where hosts tagged as medical devices have been geo-located. Most hosts have their origin in the regions of Hessen (Germany) and Madrid (Spain), where cloud datacenters are located.

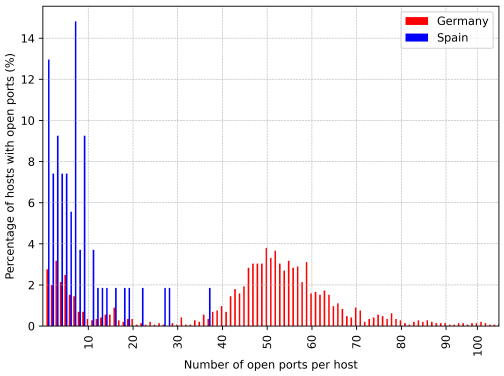


Figure 2: Percentage of hosts tagged as medical devices with a given number of open ports per host in Germany and Spain.

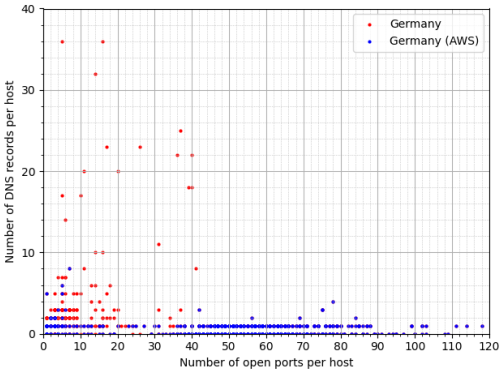


Figure 4: Correlation between the number of open ports per host and the number of DNS records per host for Germany.

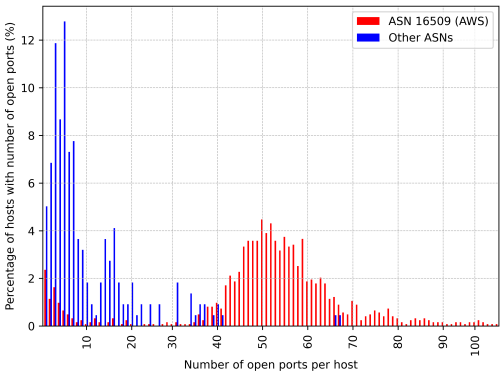


Figure 3: Percentage of German hosts tagged as medical devices with a given number of open ports per host classified by the ASN where the host IP address is originated from.

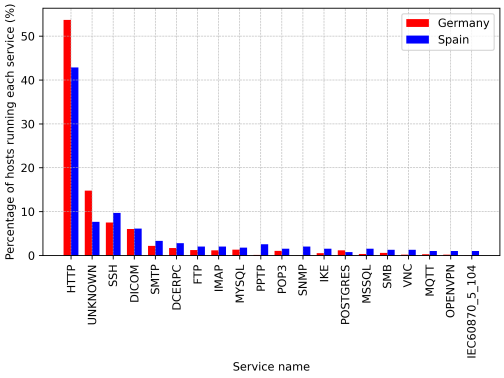


Figure 5: Percentage of most common services run by hosts tagged as medical devices in Germany and Spain.

of the hosts, whereas Windows represents 15% of the hosts. Unfortunately, for percentage larger than 50% of hosts from German and Spain the operating system could not be correctly detected. One possible explanation is that some hosts are running behind a firewall implementing destination NAT, so properly identifying the operating system of each host is not an easy task since different ports can refer to different hosts running different operating systems. Moreover, it has not been possible to collect additional information regarding the operating system of each host (i.e., version, kernel, etc.), since it is not provided by either Censys or Shodan.

We now focus on the vulnerabilities found in the services running on each host. In particular, we analyse the CVEs that affect each host/service, and the CVSS score that determines the severity of the vulnerability. This is an important aspect for hosts exposed to the Internet, since an exploitable vulnerability (i.e., buffer overflows, SQL injection, etc.) can render

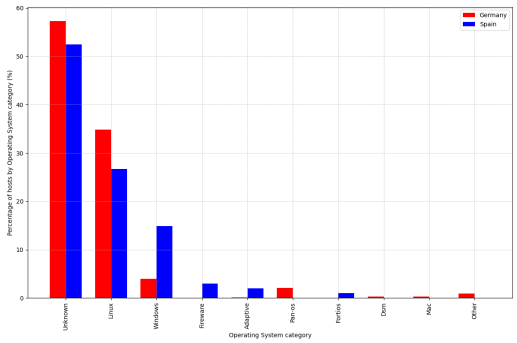


Figure 6: Percentage of most common operating systems run by hosts tagged as medical devices in Germany and Spain.

the authentication of the server useless. We do so by looking at the '`vulns`' tag provided by Shodan for each service.

Figure 7 shows the percentage of German (red) and Spanish (blue) hosts that belong to each CVSS category, with values ranging from 0 to 10 depending on the severity of the vulnerability. As it can be observed, more than 20% of German hosts and more than 15% of Spanish hosts are affected by at least one vulnerability with a CVSS value between 9.0 and 10.0, indicating a critical vulnerability that is easy for attackers to exploit and has a large compromise potential of the target system. Moreover, more than 40% of German hosts and more than 30% of Spanish hosts are affected by at least one vulnerability with a CVSS between 7.0 and 8.9 representing a vulnerability that is harder to exploit but which leads to a high compromise potential on the target.

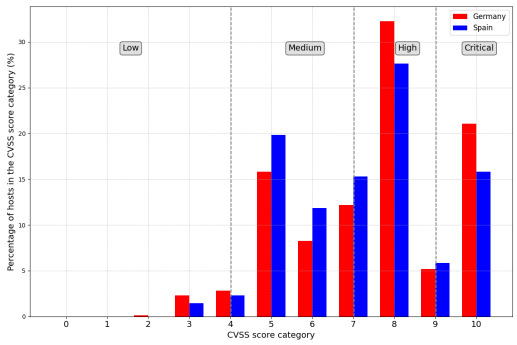


Figure 7: Percentage of hosts tagged as medical devices in German and Spain with CVEs that belong to each CVSS category (i.e., low, medium, high, and critical).

In addition, Figure 8 shows the percentage of German (red) and Spanish (blue) hosts with the top 30 CVE vulnerabilities found and their related CVSS score. As it can be observed, most of the top 30 CVE vulnerabilities found have CVSS values lower than 8, indicating that they are either difficult to exploit or that they lead to a small compromise of the target host/service. However, we also observe various CVE vulnerabilities with CVSS scores larger than 8, indicating that they are easy to exploit or that they can lead to a large compromise of the target system. These CVEs and their associate CVSS score are the following: CVE-2024-38476 (CVSS=9.8), CVE-2024-38474 (CVSS=9.8), CVE-2021-44790 (CVSS=9.8), CVE-2022-31813 (CVSS=9.8), CVE-2021-39275 (CVSS=9.8), CVE-2022-22720 (CVSS=9.8), CVE-2022-22721 (CVSS=9.1), CVE-2021-40438 (CVSS=9.0). It is worth noting that most of the CVEs found affect the Apache HTTP Server leading to DOS (Denial of Service) or RCE (Remote Code Execution).

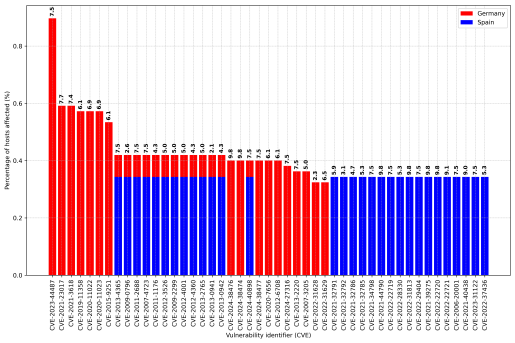


Figure 8: Percentage hosts tagged as medical devices in Germany and Spain with the top 30 vulnerabilities found and their related CVSS score.

Figure 9 shows the distribution of CVEs found in German (red) and Spanish (blue) hosts classified by the year that they were discovered and classified into the CVE system. As it can be observed, there is a large percentage of CVEs that found in these hosts that have a date In particular, the percentage of CVEs found in German hosts that date from earlier than 2020 is 74%, whereas for Spain that number is about 60%.

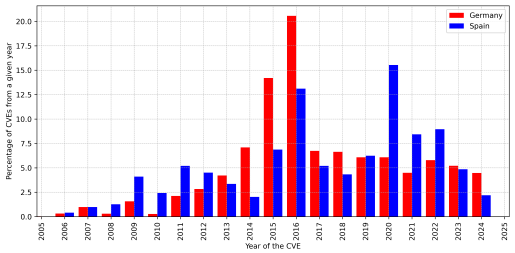


Figure 9: Percentage of CVEs of hosts tagged as medical devices in Germany and Spain classified according to the year where they were discovered.

To finalize our analysis we study the screenshots that we have obtained for each port that runs an HTTP service on each host. Using such approach, we have detected several installations with default parameters, accessible without credentials, and meant for demonstration purposes. We provide one example of each category that we have found in Germany. First, the host `62.171.155.131` runs an Orthanc installation with a insecure setup warning. Second, the host `82.165.144.45` runs an Orthanc ICQ Cloud software that is accessible without credentials. Finally, the host `212.132.115.76` is pointed by the DNS record `demo.radio.kapsiki.net`, indicating that this is not running a real medical IT service.

Analysing the remaining results we have also found several hosts that are worth describing. As a summary of the results, we only focus on three hosts running in Spain that have been tagged as medical devices. The idea is to provide a general overview of the current state of cybersecurity in hosts tagged as medical devices, allowing to discuss the threats and the potential solutions in the following subsection.

The first host `2.136.109.191` belongs to an image diagnostics centre located in Catalunya, as obtained using a reverse DNS query (i.e., `pacs.imaginebarcelona.com` domain) and the IP geo-position. Analysing the Censys results shows that the host has 4 ports open: 104 (DICOM), 161 (SNMP), 9090 (HTTP), 9091 (HTTP). In addition, analysing the Shodan results for the host shows that there are no known CVEs affecting it. Having a low number of services exposed to the Internet and no known CVEs, which is a good sign of cybersecurity hygiene. However, we are unsure why these ports are even required to be exposed to the Internet, specially DICOM and SNMP which are not meant for end-users accessing a service. Analysing the HTTP services on running ports TCP/9090 and TCP/9091 shows the landing page of a DAIPACS, a PACS software.

The second host `217.127.207.41` belongs to a medical centre located in the Castilla y Leon, as obtained using a reverse DNS query and the IP geo-position. The host has more than 10 ports publicly exposed, and most of these ports are accessible using HTTP. Through a quick analysis of the data provided by Censys and Shodan we determine that these ports expose services such an SCADA server (Tridium Niagara), a DICOM viewer (OHIF) and a database (MySQL). In addition, one of the HTTP services exposes the landing page of a firewall (Sophos UTM), indicating that most of these services run behind a NAT implementing redirection to internal IP addresses based on the destination port. Besides the large number of ports, the most relevant aspect it that the host has been flagged with more than 21 CVEs, with CVSS ranging from low to critical. In particular, the hosts contains the following CVEs (and they related CVSS) that are specially relevant: CVE-2017-8923 (CVSS=9.8), CVE-2017-9120 (CVSS=9.8), CVE-2021-21708 (CVSS=8.2), CVE-2022-31625 (CVSS=8.1), and CVE-2022-37454 (CVSS=9.8). Again, the vulnerabilities found in the host have a high CVSS score, indicating a high risk potential. Moreover, the CVEs mostly affect the Apache HTTP server or related libraries (i.e., PHP), and can lead to DOS and RCE.

The final interesting host (`83.48.110.205`) that we found in Spain belongs to a medical centre located in the Madrid region, as validated using a reverse DNS query (i.e., `traumasport.com` domain) and the IP geo-position. According to both Censys and Shodan the host runs the Microsoft Windows Server 2012 operating system and only has 6 ports open. Two ports (TCP/80 and TCP/443) run HTTP/HTTPS to host the domain website. In addition, the host also has ports TCP/137 (NetBIOS) and ports TCP/5901 and TCP/5915 (VNC) open to the public. Finally, the host also has port TCP/8999 open, which is linked to the RemotEyeLite Server v2.2.1. Despite the short number of ports, it is already concerning that NetBIOS and VNC are running on ports exposed to the Internet. However, the most eye-catching part is that, despite the short number of ports, the host is vulnerable to more than 200 CVEs. We will not describe each CVE in detail as we have done with other hosts due to the lack of space, but it is enough to say that there are several CVEs dated from 2012 (i.e., CVE-2012-4360 and CVE-2012-4001), 2011 (i.e., CVE-2011-2688 and CVE-2011-4718) and earlier, all of them with CVSS values ranging from low to high.

D. Discussion

In the previous subsection we have provided a general overview of the results obtained when querying Censys and Shodan for hosts in Germany and Spain tagged as medical devices, as well as a deeper analysis of three specific hosts located in Spain to showcase the most common vulnerabilities found. While we do not provide an in-deep analysis of hosts tagged as medical devices running in Germany due to lack of space, a quick analysis revealed that the results obtained show a similar situation to the one in Spain. Hence, in this subsection we discuss the cybersecurity outlook of medical IT devices in Germany and Spain, and provide some general guidelines to improve the situation.

1) Running unrelated services in one host: We have observed hosts tagged as medical IT devices that have multiple ports exposed and run services that are completely unrelated to their category, such as remote control (i.e., VNC) or video recording (i.e. DVR) software, among others. At this point, we should investigate further to determine if the IP address of the host uses a NAT service to redirecting connections to internal IP addresses based on the destination port. However, if it is really a system on which medical IT services are running with other unrelated services, this is a clear violation of good practice for secure IT operations.

2) Running services that are not patched: We have found hosts that are running services with known CVEs that are quite old (i.e., 10+ years) and have a high CVSS scores (i.e., 8+). Specifically, most of the vulnerabilities that we have detected affect the HTTP server handling the requests (i.e., Apache HTTP server) or related libraries (i.e., PHP) used by the applications. For example, checking the CVEs information we discovered that that the version of Apache that most hosts are running is at least 3+ years old. Hence, while the application itself could be developed according to modern security standards, a bug in the HTTP server or the related libraries may provide a malicious user a mechanism to infiltrate the server and get access to user data.

3) HTTP landing pages for medical IT services: Most medical IT services that are exposed to the Internet use the HTTP protocol to allow users connect remotely using a web browser. While most medical IT services found have a login page that requires the user to introduce a username/password to access the service, it is unknown if there are other mechanisms to protect the service behind it. For instance, there could be a default username and password if the installation uses the default configuration, as we have found in some instances through a message in the login page. Also, even if properly configured, there may not be a maximum limit of login retries or a two-factor authentication mechanism. In fact, judging by the appearance of the user interface, it seems plausible that these limitations do not exist, which is unacceptable considering modern IT security practices.

4) Re-using DNS domains for different purposes: We have observed that some medical centres reuse the domain that they use for their main website (i.e., `domain.com`) to point to hosts that run medical IT services by using a subdomain (i.e., `pacs.domain.com`). While this is correct from a technical perspective, reusing the main domain provides a potential attacker with additional information about the owner of the medical services (i.e., where it is located, who are the employees, etc.), which can be exploited to perform other

types of social engineering attacks (i.e., phishing). Hence, such practices should be avoided whenever possible.

5) Improving cybersecurity of medical IT services: As we have just described, there are many threats potentially affecting medical IT services exposed to the Internet.

Hence, to improve cybersecurity of medical IT services we believe that they should be separated from general IT services, and the former should not be accessible from the Internet when possible. Instead, a VPN (Virtual Private Network) should be used to ensure that authorized users can connect to it. The VPN can be a client-to-site VPN in case it is an end-user (i.e., a doctor) that needs to access the service, or a site-to-site VPN in case two medical IT systems have to connect to each other (i.e., a CT or MRI and a PACS). Currently, there are many open-source projects (i.e., WireGuard) that provide simple and affordable ways to deploy such tunnels, so not using them is gross misconduct in terms of cybersecurity hygiene. Moreover, in case of medical IT services that cannot be hidden behind a VPN, we would recommend to: 1) separate the services using a firewall with NAT, 2) separate the domains used for general IT services (i.e., website) from medical IT services (i.e., PACS server). Finally, the operating system and the software running the medical IT services should be frequently patched, and each service should be appropriately hardened, with a limitation on the maximum number of logins or a double-factor authentication to limit the risks of attacks.

Last but not least, we want to stress that the methods used throughout the paper did not involve any active hacking. Moreover, the OSINT tools that we have used are accessible to the general public, only requiring registration and not involving any subscription fees when used for a limited number of hosts. Hence, system/network administrators responsible for medical IT systems should use these tools to gather information about potential security holes caused by lack of updates or misconfiguration, anticipating potential threats before they occur.

V. Conclusions and future work

This study has investigated the cybersecurity status of Internet-exposed medical IT systems in Germany and Spain using a processing pipeline that gathers and analyses data from OSINT tools. We found that 20% of German hosts and 15% of Spanish hosts tagged as medical devices have at least one critical vulnerability, with CVSS equal or greater to 8. We also discovered that vulnerabilities found in these systems are old, with 74% of CVEs in German systems and 60% of CVEs in Spain dating from before 2020. Hence, to improve cybersecurity, we recommend applying common IT practices to medical IT systems, including: 1) regular updating of the operating system and the services, 2) implementing strict access control measures, such as using VPNs for remote access, 3) separating general IT services from medical IT systems to minimize attack surface, and, 4) monitoring OSINT data to proactively identify potential vulnerabilities.

Based on our findings, in the future we also plan to: 1) investigate specific vulnerabilities found in this study, providing detailed mitigation strategies and exploitation prevention techniques, 2) monitor the evolution of CVEs identified in our study over time to assess the effectiveness of remediation actions taken by administrators, and, 3) conduct a worldwide study using the developed methodology and tools to assess the cybersecurity status of Internet-facing medical IT systems across various countries. In addition, we believe it would be valuable to foster the collaboration with with medical IT vendors, healthcare providers, and regulatory bodies, to integrate our findings into their cybersecurity guidelines.

Acknowledgements

This project is carried out within the framework of the Recovery, Transformation, and Resilience Plan, funded by the EU Next Generation funds, under the auspices of the INCIBE cybersecurity chair named CARISMATICA. The authors of the project would also like to acknowledge the collaboration of Censys in the development of the project by providing unrestricted access to their tools.

References

[1] "DICOM (Digital Imaging and COmmunications in Medicine)," https://www.dicomstandard.org, Last accessed: 2025-02-17.

[2] L. Yan, "DICOM Standard and Its Application in PACS System," vol. 1, no. 1. [Online]. Available: http://systems.enpress-publisher.com/index.php/MIPT/article/view/221

[3] M. Stites and O. S. Pianykh, "How Secure Is Your Radiology Department? Mapping Digital Radiology Adoption and Security Worldwide," *American Journal of Roentgenology*, vol. 206, no. 4, pp. 797–804, 2016, pMID: 26934387. [Online]. Available: https://doi.org/10.2214/AJR.15.15283

[4] G. N. GmbH, "Unprotected patient data in the Internet." [Online]. Available: https://www.greenbone.net/wp-content/uploads/Greenbone_Security_Report_Unprotected_Patient_Data_a_Review.pdf

[5] S. Yazdanmehr. Millions of Patient Records at Risk. aplite. [Online]. Available: https://i.blackhat.com/EU-23/Presentations/EU-23-Yazdanmehr-Millions_of_Patient_Records_at_Risk.pdf

[6] "The Global State of Internet of Healthcare Things (IoHT) Exposures on Public-Facing Networks," https://censys.com/state-of-internet-of-healthcare-things/, October 2024, Last accessed: 2025-02-17.

[7] M. H. Pervez, M. I. Ecevit, N. Z. Naqvi, R. Creutzburg, and H. Dag, "Towards Better Cyber Security Consciousness: The Ease and Danger of OSINT Tools in Exposing Critical Infrastructure Vulnerabilities," in *2023 8th International Conference on Computer Science and Engineering (UBMK)*. IEEE, pp. 438–443. [Online]. Available: https://ieeexplore.ieee.org/document/10286573/

[8] S. B. Weber, S. Stein, M. Pilgermann, and T. Schrader, "Attack Detection for Medical Cyber-Physical Systems–A Systematic Literature Review," vol. 11, pp. 41 796–41 815.

[9] C. P. Research, "38% increase in 2022 global cyberattackss." [Online]. Available: https://blog.checkpoint.com/2023/01/05/38-increase-in-2022-global-cyberattacks/

[10] "ENISA threat landscape: health sector (January 2021 to March 2023)." [Online]. Available: https://www.enisa.europa.eu/publications/health-threat-landscape

[11] H. S. C. C. Center, "HC3's Top 10 Most Active Ransomware Groups. Analyst Note 202404051700." [Online]. Available: https://www.aha.org/system/files/media/file/2024/04/HC3-tlp-clear-analyst-note-top-10-most-active-ransomware-groups-4-10-24.pdf

[12] "ENISA threat landscape 2024: July 2023 to june 2024." [Online]. Available: https://www.enisa.europa.eu/sites/default/files/2024-11/ENISA%20Threat%20Landscape%202024_0.pdf

[13] M. Walkowski, J. Oko, and S. Sujecki, "Vulnerability Management Models Using a Common Vulnerability Scoring System," vol. 11, no. 18, p. 8735. [Online]. Available: https://www.mdpi.com/2076-3417/11/18/8735

[14] H. Berghel, "Robert David Steele on OSINT," vol. 47, no. 7, pp. 76–81. [Online]. Available: https://ieeexplore.ieee.org/document/6861897/

[15] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast Internet-wide Scanning and Its Security Applications," in *22nd USENIX Security Symposium (USENIX Security 13)*. Washington, D.C.: USENIX Association, Aug. 2013, pp. 605–620. [Online]. Available: https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/durumeric

[16] M. Aggarwal, "A Study of CVSS v4.0: A CVE Scoring System," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, pp. 1180–1186, place: Gautam Buddha Nagar, India. [Online]. Available: https://ieeexplore.ieee.org/document/10397701/

[17] P. Heinl, A. Patapovas, and M. Pilgermann, "Towards AI-enabled Cyber Threat Assessment in the Health Sector," version Number: 1. [Online]. Available: https://arxiv.org/abs/2409.12765

From Theory to Practice: Assessing the Suitability of Testbeds in Industrial Cybersecurity

Jovan A. Guillén-Mass Rafael A. Rodríguez-Gómez Roberto Magán-Carrión
Network Engineering & Security Group (NESG)
Research Centre for Information and Communication Technologies (CITIC-UGR)
University of Granada, Granada, Spain.
jaguillenmass@correo.ugr.es, {rodgom, rmagan}@ugr.es

***Abstract*—Cybersecurity in industrial environments is critical for safeguarding essential infrastructures such as power plants and factories. The increasing convergence of Operational Technology (OT) with Information Technology (IT) has introduced new vulnerabilities, amplifying the risks faced by industrial control systems (ICS). While IT systems have more mature cybersecurity frameworks, OT environments face unique challenges like outdated systems, insufficient security updates, and limited threat visibility. Achieving cybersecurity maturity in OT requires not only advanced technological tools but also specialized testing environments to simulate and assess potential threats without impacting real systems. This paper explores the approaches and trends used in the construction of industrial control system testbeds over the past five years evaluating their capabilities, and explores the future potential of these technologies, emphasizing the need for research and innovation in this area.**

***Index Terms*—Cybersecurity, Industrial Control Systems (ICS), Operational Technology (OT), Testbeds, Virtualization, Containerization, SDN.**

Contribution type: *Ongoing investigation*

I. Introduction

Cybersecurity in industrial environments is a key factor for protecting critical infrastructures such as power plants, factories, and water distribution networks. These operational technology (OT) systems have traditionally been isolated from IT networks, but the growing convergence of both domains has created new vulnerabilities and risks that must be properly managed. According to Fortinet [1], 31 percent of OT organizations reported six or more intrusions in 2024, highlighting a significant increase in attacks, with phishing and business email compromise as the most common threats (58 percent). Moreover, only 5 percent of companies have full visibility over their OT environment, emphasizing the need for better security governance and automation.

The path to cybersecurity maturity in industrial environments involves not only adopting technological tools but also developing capabilities to investigate, test, and simulate cybersecurity scenarios without compromising the operation of real systems. The lack of adequate testing environments in OT is one of the biggest obstacles, as performing tests directly on operational systems can be risky. In addition, attacks targeting industrial environments are becoming more sophisticated, making research and simulations vital to anticipating and mitigating future incidents.

In this context, the research and development of specialized testing environments emerge as an urgent need to achieve effective cybersecurity in industrial settings. Conti et Al. [2] provide a comprehensive review of existing testbeds and datasets for industrial control system (ICS) security research, highlighting key trends, challenges, and gaps. It emphasizes the necessity of realistic and well-documented datasets to enhance threat detection and cybersecurity analysis. Similarly, Di Modica and Foschini [3] examine the evolution of ICS testbeds, classifying them based on their design approaches and identifying critical aspects for improving their fidelity and scalability. Both works underline the importance of developing robust, flexible, and representative test environments to advance industrial cybersecurity research.

This paper complements the previously mentioned papers by providing an analysis of several studies in which industrial control system (ICS) testbeds were created to test hypotheses, evaluate prototypes or cybersecurity solutions, and measure system behaviors. It examines the technologies applied in these testbeds and, through a subjective assessment, explores the potential of each technology to expand the testing environment. The goal is to identify trends and determine the ideal characteristics a testbed should have to ensure high fidelity, flexibility, scalability, automation capabilities, and the ability to simulate various cyberattacks.

Identified trends in ICS testbeds highlight the widespread use of virtualization for scalability, the importance of network emulation for realistic scenarios, and the preference for simulated physical processes to enhance data quality. While physical testbeds offer higher fidelity, virtualized environments are more cost-effective. Additionally, dedicated images for industrial components and specialized attack instances are prioritized for security research.

This paper is structured as follows: Section II pro-

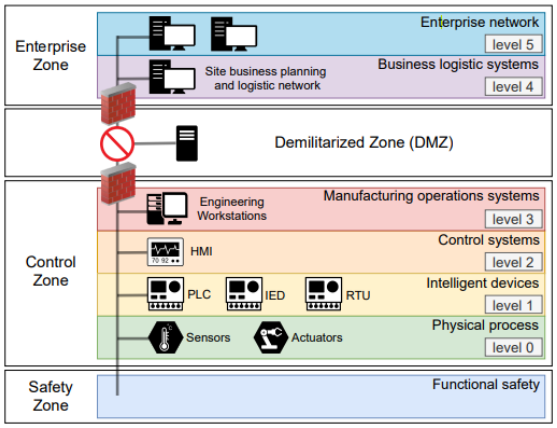


Figure 1. The Purdue Model (Conti et al. [2]).

vides a theoretical foundation, detailing ICS components, virtualization methods and technologies, and key cybersecurity concepts. Next, Section III introduces the followed methodology for analyzing the selected works. Later, in Section IV, we performed the evaluation of such chosen works which represents the current state-of-the-art research on cybersecurity testbeds for OT, assessing their current and future potential. After this, in Section IV-A a thorough discussion will be performed, identifying key trends and initial characteristics observed in the reviewed testbeds. Finally, in Section V the conclusions and emerging trends are highlighted as well as, what we think are the key requirements for building ICS testbeds and the future directions for enhancing cybersecurity testing in industrial environments.

II. Background

This section has been conceived to introduce and describe the key actors, environments and technologies for an easy understanding and following of the rest of the work. Among others, the Industrial Control Systems (ICS) will be introduced as they are target environment of the study. Furthermore, components, methodologies for simulation, virtualization and containerization that will facilitate the generation of robust datasets to be afterwards used for many different aims, will be introduced as well.

In this regards, we can distinguish two main subsections: ICS and Virtualization-related technologies.

A. Industrial Control Systems

Figure 1 shows and standard architecture of a ICS systems following the Purdue Model approach[2]. In the following, a detailed description of the involved elements shown in this model will be introduced.

- Industrial processes: series of operations performed in the making, treatment or transportation of a product or material [4].
- Industrial Internet of things: A system comprising networked smart objects, cyber-physical assets, associated generic information technologies and optional cloud or edge computing platforms, which enable real-time, intelligent, and autonomous access, collection, analysis, communications, and exchange of process, product and/or service information, within the industrial environment, so as to optimise overall production value [5].
- Industrial control systems: collection of personnel, hardware, and software that can affect or influence the safe, secure, and reliable operation of an industrial process [4].
- Programmable Logic Controller (PLC): is a microprocessor-controlled electronic device that reads input signals from sensors, executes programmed instructions using these inputs and orders from supervisory controllers, and creates output signals that may change switch settings or move actuators [2].
- Human-Machine Interface (HMI): is a software installed on desktop computers, tablets, smartphones, or dedicated flat panel screens that permit operators to check and monitor the automation processes [2].
- Supervisory Control and Data Acquisition (SCADA): is a type of loosely coupled distributed monitoring and control system commonly associated with electric power transmission and distribution systems, oil and gas pipelines, and water and sewage systems [4].
- Engineering Workstation: is generally a desktop computer or server running a standard operating system hosting the various software for controller and applications. Engineers use this platform to manage the controllers [2].
- Data Historian: a software application used to collect real-time data from the processes and aggregate them into a database for analysis [2].
- Remote Terminal Unit (RTU): A RTU is a microprocessor-controlled electronic device. Like PLC, it is designed for harsh environments and is generally located far from the control center, for instance, in voltage switch-gear [2].
- Intelligent Electronic Device (IED): is a device containing one or more processors that can receive or send data from an external source. Examples of IEDs are electronic multi-function meters, digital relays, and controllers [2].
- Network: arrangement of hardware, software, and propagation media to allow the transfer of messages from one application to another [4].
- Industrial network protocols: protocols that have been developed to support the specific requirements of the OT environment, like fault tolerance and reliability [2]. Examples include Modbus, PROFINET, EtherNet/IP, and OPC UA.
- Modbus: a serial communication protocol initially published by Modicon (now Schneider Electric)

in 1979 to use with its programmable logic controllers (PLCs) [2].
- Profinet: This protocol follows the provider-consumer model for data exchange in a cascading real-time concept. It is compatible with Ethernet thanks to its flexible line, ring, star structures, and copper and fiber-optic cable solutions [2].
- EtherNet/IP: is an adaption of the Common Industrial Protocol (CIP) to the Ethernet TCP/IP stack [2] .
- OPC Unified Architecture (OPC UA): is an operational framework communications in process control systems.This specification is built around Service-Oriented Architecture (SOA) and is based on web services, making it easier to implement OPC connections over the Internet [2].
- Demilitarized Zone (DMZ): zone where of data exchange between the OT environment and the IT environment, managing the connection between them [2].
- Enterprise Resource Planning (ERP): software platforms that are designed to unify and manage an organization's core business processes, such as finance, human resources, supply chain, and manufacturing. In the context of industrial control systems, ERP systems facilitate the seamless integration of operational technology (OT) and information technology (IT), enhancing data flow and coordination across various departments. This integration enables real-time monitoring and control, leading to improved decision-making and operational efficiency.
- Firewall: is a network security device designed to monitor, filter, and control incoming and outgoing network traffic based on predetermined security rules. The primary purpose of a firewall is to establish a barrier between a trusted internal network and untrusted external networks [6].
- Safety Zone: this is where devices and systems are deployed for managing ICS security by monitoring for anomalies and avoiding dangerous failure [2].
- Field Devices: include all the components that are in direct contact with the physical process.The controllers can use them to get information regarding the physical process (e.g., the measure of temperature or pressure using sensors). Instead, actuators can interact with a physical process following commands from a controller (e.g., control motors, pumps, valves, turbines, agitators) [2].

B. Virtualization-related technologies

- Testbed: scaled-down versions of a real ICS, created ad-hoc to reproduce real-world systems but in a controlled environment [2].
- Virtualization: Is a technology that has been widely applied for sharing the capabilities of physical computers by splitting the resources among OSs [7]. This helps to create virtual computing environments (e.g., virtual machines) to optimize hardware resource utilization, improve scalability, and enhance security in industrial systems.
- Containerization: type of virtualization technology that is capable of encapsulating functionalities into modules with customizable granularity. In contrast to monolithic architectures, the lightweight containerization paradigm fosters the adoption of finer-grained modularized components, each specialized in a specifc task, enforcing the loose coupling/high cohesion programming principles [3].
- Kubernetes (K8S): is the de-facto standard for container orchestration. i.e. installing a cluster of worker Nodes and deploying applications on these worker Nodes as a set of Pods, which are the unit of deployment in K8s. A single Pod encapsulates one or more containers that logically belong together. Network Policies in K8s allow specifying firewall rules on a per Pod basis to restrict communication with other Pods according to the principle of least privilege[8].
- Software-Defined Networking (SDN): is an approach to design, implement and manage networks that separate the network control (control plane) for a better user experience. This network segmentation offers numerous benefits in terms of flexibility and control [9].

III. Methodology

In order to build a solid understanding of existing approaches to ICS testbeds, we have selected 16 research papers published within the last five years in highly relevant international journals. For that, databases like Scopus, IEEE explore and the Web of Knowledge has been consulted. While the primary objectives of these studies varied—ranging from hypothesis testing to evaluating specific products, behaviors, or security mechanisms—all of them implemented some form of an ICS testbed. Moreover, these studies provide valuable insights into different methodologies, architectures, and technologies used in testbed development.

The analysis begins with a classification of the selected papers based on their degree of virtualization: **virtualized, hybrid, or physical** implementations. Subsequently, an evaluation framework is introduced, highlighting key criteria for assessing the effectiveness, scalability, and realism of these testbeds together with a simple quantitative analysis. Finally, a comparative analysis is presented, summarizing the results, identifying trends, strengths and limitations of the studied works.

A. Classification

The testbeds are classified into three main categories based on their degree of virtualization: virtualized,

hybrid, and physical. This classification is essential for identifying the most suitable testbed approach for specific use cases, balancing between high-fidelity physical interactions and cost-effective simulations.

Fidelity, defined as the accuracy with which a testbed replicates real ICS systems, is a critical factor, often linked to higher costs. Physical testbeds provide the most accurate simulations but are costly, while virtualized testbeds are more affordable yet may compromise on realism. However, advancements in technologies like Software-Defined Networking (SDN) and the use of mathematical models in tools such as MATLAB or Python can enhance the fidelity of virtualized environments. Additionally, hybrid approaches that integrate virtual machines and containerized elements such as PLCs, HMIs, and SCADA systems offer a promising balance between accuracy and cost.

This classification framework is intended to guide the selection of testbed designs for ICS experiments. A detailed analysis of each category is provided in the following sections.

- Physical Testbeds: provide high fidelity by replicating real-world systems and their underlying hardware, making them ideal for accurate simulations of industrial processes and real-time interactions. However, they are costly due to the need for dedicated infrastructure.
- Virtualized Testbeds: are more cost-effective, using virtual machines or containers to replicate environments, though they offer lower fidelity since they cannot fully replicate physical interactions. However, they are easily scalable and suited for testing software, protocols, and algorithms, with technologies like SDN enhancing network simulation.
- Hybrid Testbeds: combine both virtual and physical components, offering a balance between cost and fidelity by maintaining key hardware elements while virtualizing others, such as networks and controllers. Consequently, this hybrid approach provides flexibility and is useful when certain real-world interactions need to be preserved within cost-related constraints.

Each type of testbed has its own advantages and drawbacks, and selecting the right configuration depends on the specific goals of the research or test. By understanding these differences, it is possible to build an effective and efficient testbed that aligns with both cost and technical requirements.

B. Evaluation criteria

The evaluation is primarily objective, based on the information provided by the authors. However, drawing on our expertise, we also assess the potential capabilities of the technologies used, offering a complementary perspective beyond what is explicitly stated.

For each criteria a High (●), Medium (●) and Low (●) degree of achievement is considered for its evaluation with different meaning depending on the specific criteria.

1) Scalability: The ability of the testbed environment to grow in size and complexity without compromising performance or stability is known as scalability. This includes the number of system elements (PLCs, sensors, actuators, servers, etc.) it can support, as well as the variety of compatible industrial connectivity protocols. Scalability testbed criteria degrees are the following:

- High: The ability of generation of architectures with more than 20 elements and supports at least 5 industrial connectivity protocols (e.g., Modbus, OPC UA, Profinet, EtherNet/IP, DNP3, etc.).
- Medium: The ability of generation of supporting between 10 and 20 elements and at least 3 industrial connectivity protocols.
- Low: 10 or fewer elements and supports 2 or fewer industrial connectivity protocols is considered as the lowest scalability of a testbed.

2) Integration: The ability of the testbed environment to connect with other third-party systems, tools or technologies. Integration criteria degrees are described ahead:

- High: The ability to connect with more than 5 external technologies without the need for complex modifications (analysis platforms, simulators, sensors, etc.).
- Medium: The ability to connect between 3 and 5 external technologies, but requires certain adjustments or additional configurations.
- Low: the lowest level implies that only 2 or fewer external technologies con be connected, with a high dependency on manual configurations or adaptations.

3) Flexibility: The degree to which the testbed allows for the configuration and modification of components, protocols, or architectures without requiring significant changes to its structure. Flexibility testbed criteria degrees are the following:

- High: The testbed allows extensive modifications in architecture, protocols, and configuration with ease, adapting to various use cases.
- Medium: The testbed supports certain changes, but with restrictions or significant manual intervention required.
- Low: the lowest level implies rigid configuration with limited options for modification and adaptation.

4) Automation Capability: The degree to which the installation, configuration, and maintenance of the testbed can be automated, minimizing the need for manual intervention by specialized personnel. This includes the use of scripts, orchestration tools, and automated workflows to streamline processes and reduce

the likelihood of human error. Automation criteria degrees are described ahead:

- High: The testbed can be Fully automated deployment using scripts, infrastructure as code (IaC), or orchestration tools like Ansible or Docker Compose.
- Medium: The testbed requires manual intervention in some steps, but has some degree of automation in the installation or initial configuration.
- Low: The installation and configuration are mostly manual within the testbed, with little to no automation in implementation.

5) Attack Simulation Capability: The ability to simulate realistic cyberattacks within the testbed, covering various tactics, techniques, and procedures used in industrial attacks. Attack simulation testbed criteria degrees are the following:

- High: The testbed allows for the execution of automated and customizable attacks with advanced tools (e.g., Metasploit, Scapy, Kali Linux) and supports complex scenarios.
- Medium: The tesbed Supports predefined attacks, but with limitations in customization or automation.
- Low: The lowest level implies that attack simulation is limited or nonexistent, with only manual tests or very specific cases.

6) Dataset Generation Capability: : The ease with which the testbed can collect and store data on events, behaviors, and anomalies for a later analysis and use in security models or algorithms. Dataset generation testbed criteria degrees are described ahead:

- High: The testbed can generate large volumes of labeled data, with logging tools, export in multiple formats, and compatibility with databases.
- Medium: Dataset generation is possible within the testbed, but with limited labeling or storage options.
- Low: When the testbed is limited or has no capability to generate structured datasets for analysis falls into the lowest degree.

This evaluation criteria will provide a clear understanding of how each testbed performs across the different dimensions outlined, highlighting their strengths and limitations. These results will serve as the foundation for the subsequent quantitative analysis which is described next.

C. Quantitative Analysis

In this phase, a scoring system was applied based on the evaluation criteria outlined in the previous subsection. Each criterion was assigned a value of Low = 1, Medium = 2, and High = 3, allowing for the calculation of an average score. This approach was designed to identify which configurations achieved the best and worst average scores within their respective classifications. By analyzing these results, it was possible to determine the configurations that most effectively met the research goals.

IV. Comparative analysis

Table I provides a comprehensive summary of the 16 studied testbeds for Industrial Control Systems (ICS) cybersecurity. They have been organized by publication year and classified by virtualization type. The rest of the columns in this table, show the previous introduced evaluation criteria and the degree of achievement both from an objective and subjective perspective.

As it is shown in Table I, most of the testbeds are virtualized, with 10 out of 16 relying primarily on containerization technologies (e.g. Docker) to emulate key ICS components including PLCs, SCADA systems, and process simulators. Four of them adopt an hybrid approach by integrating physical hardware alongside virtual environments, while only two are entirely physical. The diverse array of technologies used, such as OpenPLC, SCADABR, FUXA, Python-based simulations, and network emulation tools like GNS3, NS3, and Mininet, underscores the multifaceted nature of ICS Testbed construction. Moreover, some testbeds incorporate advanced solutions like SDN and dedicated virtual machines running, for instance, a Kali Linux instance to simulate sophisticated cyberattack scenarios and Mathlab or python based physical process simulation to increase data quality, this means ensuring that data is accurate, consistent, complete, and reliable throughout its lifecycle, with a focus on preparing it to train cybersecurity models and algorithms.

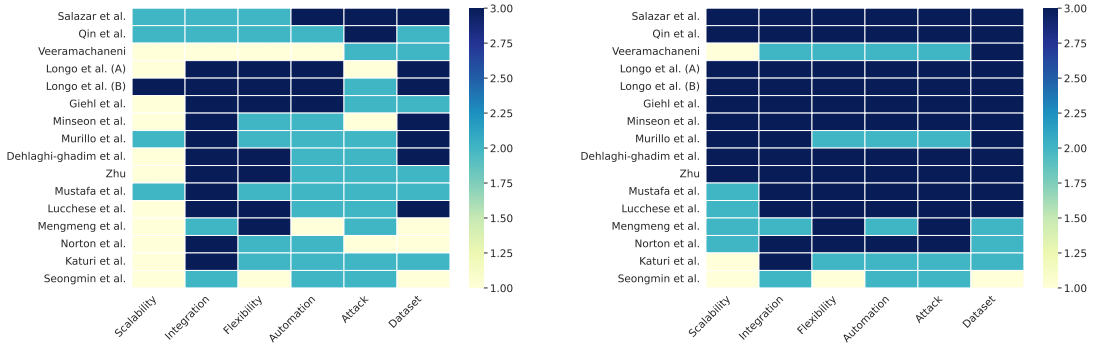
According to the evaluation metrics, they reveal significant variations depending on the previous classification. As it is shown in Table I, virtualized testbeds typically score higher in scalability and automation than the others, due to their container-based architectures and the ease of quick reconfiguration. Conversely, testbeds with physical components tend to offer greater realism but often face challenges in scalability and maintenance. In terms of integration and flexibility most testbeds are generally robust, as many effectively combine multiple technologies to create comprehensive and adaptable testing environments. Meanwhile, the capability to simulate attacks and generate structured, labeled datasets—critical for training cybersecurity models and algorithms—varies widely.

Regarding the results of the quantitative analysis, which are shown in Figure 2, some studies can be highlighted. Longo et al. [14] achieved the highest average score of 2.91 in the virtualized testbed classification, showing its robust performance in scalability, integration, and automation. Mustafa et al. [20] and Lucchese et al. [21] scored 2.41, bringing great integration capabilities. Lastly, Katuri et al. [24] highlighting its limitations in key areas like scalability but also great integration. Other examples like Minseo et

Classification	Papers	Year	AVG	Scalability		Integration		Flexibility		Automation		Attack		Dataset	
				Objective	Potential	Objective	Potential	Objective	Potential	Objective	Potential	Objective	Pontential	Objective	Potential
Fully Virtualized	Salazar et al. [10]	2024	2.75	●	●	●	●	●	●	●	●	●	●	●	●
	Quin et al. [11]	2024	2.58	●	●	●	●	●	●	●	●	●	●	●	●
	Veeramachaneni [12]	2024	1.66	●	●	●	●	●	●	●	●	●	●	●	●
	Longo et al. [13]	2024	2.66	●	●	●	●	●	●	●	●	●	●	●	●
	Longo et al. [14]	2023	2.91	●	●	●	●	●	●	●	●	●	●	●	●
	Giehl et al. [15]	2024	2.66	●	●	●	●	●	●	●	●	●	●	●	●
	Minseo et al. [16]	2023	2.58	●	●	●	●	●	●	●	●	●	●	●	●
	Murillo et al. [17]	2024	2.58	●	●	●	●	●	●	●	●	●	●	●	●
	Dehlaghi-Ghadim et al. [18]	2024	2.75	●	●	●	●	●	●	●	●	●	●	●	●
	Zhu [19]	2024	2.66	●	●	●	●	●	●	●	●	●	●	●	●
Hybrid	**Mustafa et al. [20]**	2025	2.41	●	●	●	●	●	●	●	●	●	●	●	●
	Lucchese et al. [21]	2024	2.41	●	●	●	●	●	●	●	●	●	●	●	●
	Mengmeng et al. [22]	202	2	●	●	●	●	●	●	●	●	●	●	●	●
	Norton et al. [23]	2025	2.08	●	●	●	●	●	●	●	●	●	●	●	●
Physical	**Katuri et al. [24]**	2024	2	●	●	●	●	●	●	●	●	●	●	●	●
	Seongmin et al. [25]	2025	1.33	●	●	●	●	●	●	●	●	●	●	●	●

* **Criteria level**: High (●), Medium (●) and Low (●)

Table I
COMPARATIVE ANALYSIS OF THE STUDIED SOLUTIONS FROM A VARIETY OF CRITERIA.



(a) Objective approach

(b) Potential capabilities

Figure 2. Quantitative analysis of the studied works vs the objective- and Potential Capabilities-based approach. Each work is ranging from 1 to 3 being the last the higher achievable score.

al. [16] and Murillo et al. [17] falls into the median performance which was 2.58, providing a baseline for typical capabilities in current ICS cybersecurity research. Additionally, per-classification best scored papers are highlighted in Table I in gray.

These results illustrate the diverse approaches and trade-offs in ICS testbed design. The findings emphasize that virtualized environments, often enhanced with advanced network emulation and automated deployment techniques, can offer significant benefits in scalability and ease of use, while hybrid and physical testbeds provide higher fidelity but at a greater cost and complexity. In Section IV-A, these findings will be analyzed and explored in detail as well as their implications for future ICS testbed design and development.

A. Discussion

As stated before, testbeds for Industrial Control Systems (ICS) cybersecurity are of upmost importance for testing and developing cybersecurity solutions in this context. The present study is aimed at revealing several important trends and insights that are critical for advancing both research and practical applications in this domain through the revision of recent works.

To begin with, the study reveals a clear evolution in their approaches and employed technologies with the studied period. For instance, in the early years of the analyzed period (2021-2022), the use of virtualized testbeds was mainly observed, heavily relying on virtual machines (VMs) and containerization technologies (e.g. Docker) for emulating ICS components. However, from 2023 onward, there has been a noticeable shift towards hybrid (semi-virtualized) approaches, integrating physical hardware such as real PLCs and network devices alongside virtual environments. This trend reflects an increasing need to capture the complexities of real-world industrial systems, leading to a rise in the adoption of technologies like Software-Defined Networking (SDN) and industrial traffic simulators. Additionally, the use of dedicated attack simulation environments, such as full-fledged Kali Linux instances, has grown in recent years, highlighting a stronger emphasis on active threat evaluation within these testbeds.

Building on this trend, the analysis of scalability, integration, and flexibility across the reviewed testbeds highlights the trade-offs inherent in different technological approaches. For virtualized testbeds, while offering unmatched scalability and rapid reconfiguration capabilities, often struggle with integration fidelity, as purely software-based emulations may fail to capture the physical constraints and real-time behaviors of ICS environments. In contrast, hybrid testbeds achieve better integration by incorporating physical PLCs, HMIs, and industrial network devices, ensuring a higher degree of realism. However, this

comes at the cost of reduced flexibility, as hardware dependencies limit the speed of modifications and expansions. Furthermore, the inclusion of SDN has proven to be a key factor in improving flexibility, as it enables dynamic network reconfiguration without requiring physical changes. Moreover, testbeds with dedicated attack simulation environments, such as full Kali Linux instances, exhibit greater adaptability for cybersecurity research, allowing for the execution of complex attack scenarios and the customization of penetration testing tools. These findings underscore that an optimal testbed design must balance scalability, integration, and flexibility by strategically combining virtualization, hardware components, and advanced network management technologies.

In regards to the virtualization and integration, our review highlights the growing importance of automation in ICS testbeds, with higher-scoring testbeds demonstrating extensive use of automation for deployment, configuration and orchestration. Across the analyzed papers, testbeds that incorporate some kind of automation oriented technologies—consistently achieve better scores in scalability, flexibility, and reproducibility. Di modica and Foschini [3] emphasize how microservices-based architectures, combined with automated provisioning, enable modularity and rapid adaptation to evolving security requirements. Automation could play a crucial role in enforcing security policies aligned with IEC-62443 and NERC CIP standards, reducing the likelihood of miss-configurations and human errors.

Another critical aspect is the capability to execute and customize cyberattacks effectively. Our review indicates that testbeds with dedicated attack modules—particularly those leveraging full Kali Linux virtual machines—consistently outperform those using lightweight alternatives such as Kali Linux containers or generic Linux terminals. For example, authors in [14] employ a dedicated Kali Linux VM to facilitate comprehensive attack emulation, enabling researchers to automate and execute a wide range of exploits, including advanced persistent threats (APTs) and multi-stage attacks. A full-fledged VM provides a persistent and fully customizable environment where security researchers can install and configure a wide range of tools, automate complex attack sequences, and maintain exploit databases without containerization constraints. In contrast, while a Kali Linux container offers rapid deployment and resource efficiency, it often lacks the flexibility needed for in-depth adversarial emulation, especially when dealing with kernel-level attacks or persistent threats. Generic Linux pods, though useful for basic network probing, are the least capable, as they require extensive manual setup to match the capabilities of a pre-configured offensive security platform. The benefits of a dedicated attack environment extend beyond testbed functionality—by enabling the execution of automated and customizable attack scenarios, they facilitate the generation of structured and labeled datasets. This is particularly advantageous for future research in machine learning-based intrusion detection systems, where high-quality, well-labeled data is essential for training models capable of recognizing attack patterns in real-world ICS environments.

The reviewed setups enable the collection of structured and labeled attack data, which is essential for training machine learning models. To enhance data management, integrating technologies such as SIEM, InfluxDB, Telegraf, and Chronograf is vital. These tools facilitate the storage, monitoring, and processing of large volumes of real-time data generated by the testbed, allowing for better data extraction and analysis. A well-designed testbed that incorporates these technologies not only ensures efficient data handling but also guarantees the generation of high-quality datasets. This is crucial for training generative AI modules as Large Language Models (LLM) that can recognize complex attack patterns and anomalies within ICS environments. By including SIEM for centralized event logging and InfluxDB with Telegraf for time-series data collection and visualization, the testbed can process and store the necessary datasets to improve machine learning models and bolster cybersecurity defenses.

V. Conclusions

This study reviewed sixteen recent papers on testbeds for Industrial Control Systems (ICS) cybersecurity, published between 2021 and 2025. The testbeds analyzed included fully virtualized environments using Docker, semi-virtualized setups, and fully physical testbeds. The key evaluation metrics were scalability, integration, flexibility, automation, attack simulation, and dataset generation.

We found that virtualization is crucial in modern ICS cybersecurity testbeds. Fully virtualized testbeds offer scalability and automation but may lack physical fidelity. Semi-virtualized setups balance realism with cost-effectiveness. Many testbeds also integrate attack modules and generate structured datasets for machine learning in intrusion detection.

The research recommends that future testbed configurations adopt a hybrid approach, combining virtualization with physical components to enhance fidelity without incurring high costs. The integration of Infrastructure as Code (IaC) practices and advanced network emulation tools is also essential.

In conclusion, the hybrid approach combining virtual and physical elements is most promising. Future work should focus on implementing integrated testbeds aligned with industry standards like IEC-62443, NERC CIP, NIST 800, and ISO 27001 to advance cybersecurity research and generate high-quality datasets.

ACKNOWLEDGMENTS

This work was supported by the project NetSEA-GPT (C-ING-300-UGR23) funded by Consejería de Universidad, Investigación e Innovación and the European Union through the ERDF Andalusia Program 2021-2027 and the "Recovery, Transformation, and Resilience Plan" under agreement C025/24 between INCIBE and the University of Granada, funded by the European Union (Next Generation funds).

REFERENCES

[1] Fortinet, "State of Operational Technology and Cybersecurity Report," Tech. Rep., 2024. [Online]. Available: https://www.fortinet.com/content/dam/fortinet/assets/reports/report-state-ot-cybersecurity.pdf

[2] M. Conti, D. Donadel, and F. Turrin, "A Survey on Industrial Control System Testbeds and Datasets for Security Research," *IEEE Communications Surveys & Tutorials*, 2021.

[3] G. Di Modica and L. Foschini, "A Survey on the Use of Lightweight Virtualization in I4.0 Manufacturing Environments," *Journal of Network and Systems Management*, Feb. 2023.

[4] IEC, "IEC/TS 62443-1-1," 2009.

[5] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, "The industrial internet of things (iiot): An analysis framework," *Computers in Industry*, 2018.

[6] Fortinet, "What Is A Firewall?" 2025. [Online]. Available: https://www.fortinet.com/resources/cyberglossary/firewall

[7] F. Rodríguez-Haro, F. Freitag, L. Navarro, E. Hernánchez-sánchez, N. Farías-Mendoza, J. A. Guerrero-Ibáñez, and A. González-Potes, "A summary of virtualization techniques," *Procedia Technology*, 2012.

[8] G. Budigiri, C. Baumann, E. Truyen, and W. Joosen, "Elastic cross-layer orchestration of network policies in the kubernetes stack," *IEEE Transactions on Network and Service Management*, 2025.

[9] K. Benzekki, A. El Fergougui, and A. Elbelrhiti Elalaoui, "Software-defined networking (sdn): a survey," *Security and Communication Networks*, 2016.

[10] L. Salazar, E. Lopez-Morales, J. Lozano, C. Rubio-Medrano, and A. Cárdenas, "ICSNet: A Hybrid-Interaction Honeynet for Industrial Control Systems," *Association for Computing Machinery*, 2024.

[11] X. Qin, F. Jiang, C. Dong, and R. Doss, "A hybrid cyber defense framework for reconnaissance attack in industrial control systems," *Computers Security*, 2023.

[12] V. Veeramachaneni, "Emerging Authentication Technologies for Zero Trust in IoT Systems," *Journal of Advance Research in Mobile Computing*, 2024.

[13] G. Longo, F. Lupia, A. Merlo, F. PAgano, and E. Russo, "A data anonymization methodology for security operations centers: Balancing data protection and security in industrial systems," *Information Sciences*, 2024.

[14] G. Longo, A. Orlich, S. Musante, A. Merlo, and E. Russo, "MaCySTe: A virtual testbed for maritime cybersecurity," *SoftwareX)*, 2023.

[15] A. Giehl, M. P. Heinl, and V. Embacher, "EmuFlex: A Flexible OT Testbed for Security Experiments with OPC UA," *Association for Computing Machinery*, 2024.

[16] K. Minseo, J. Seungho, J. Cho, and G. Seonghyeon, "Data-Driven ICS Network Simulation for Synthetic Data Generation," *Electronic*, 2023.

[17] A. Murillo, R. Taormina, N. Tippenhauer, and S. Galelli, "A Thorough Cybersecurity Dataset for Intrusion Detection in Smart Water Networks," *International Join Conference on Water Distribution System Analysis (WDSA) Computing and Control in the Water Industry (CCWI)*, 2024.

[18] A. Dehlaghi-Ghadim, N. Ericsson, L.-G. Magnusson, M. Eriksson, M. Helali Moghadam, A. Balador, and H. Hansson, "Using Decision Support to Fortify Industrial Control System against Cyberattacks," *International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2024.

[19] y. Zhu, "A Sandbox for Industrial Control Systems (security)," *Université catholique de Louvain*, 2024.

[20] M. Mustafa, N. Patari, S. Basumallik, and A. Srivastava, "Augmenting Decision-Making of Human-in-the-loop Operators for Resilient Cyber-Power Systems," *Institute of Electrical and Electronics Engineers (IEEE)*, 2025.

[21] M. Lucchese, G. Salerno, and A. Pugliese, "A Digital Twin-Based Approach for Detecting Cyber–Physical Attacks in ICS Using Knowledge Discovery," *Applied Sciences*, 2024.

[22] G. Mengmeng, C. Jin-Hee, K. Dongseong, D. Gaurav, and C. Ing-Ray, "Proactive Defense for Internet-of-things: Moving Target Defense With Cyberdeception," *Arxiv*, 2020.

[23] G. Norton, J. Landsborough, E. Orozco, and J. Fields, "Containerized Cozenage: Exploring the Effectiveness of High Interaction ICS Honeypot Containers," *Hawaii International Conference on System Sciences (HICSS)*, 2025.

[24] K. Katuri, S.-Y. Park, F. Miao, and J. Zhao, "Demonstration of DoS Attacks on Modbus Protocol Using an Experimental CPS Testbed," *North American Power Symposium (NAPS)*, 2024.

[25] L. Seongmin, N. Yonghun, C. Juwon, J. Guckhyeon, l. Jeongmin, J. Sanghyun, F. Abdulaziz Alfouzan, and K. Kim, "Securing Maritime Autonomous Surface Ships: Cyber Threat Scenarios and Testbed Validation," *Institute of Electrical and Electronics Engineers (IEEE)*, 2025.

Un módulo de explicabilidad basado en los valores Shapley para frameworks de detección de anomalías sobre datos secuenciales: Temporal SHAP

Manuel Franco de la Peña, Ángel Luis Perales Gómez, Lorenzo Fernández Maimó
Departamento de Ingeniería y Tecnología de Computadores
Universidad of Murcia, Espinardo, Murcia, Spain
{manuel.francop, angelluis.perales, lfmaimo}@um.es

***Resumen*—La detección de anomalías en entornos industriales críticos exige explicaciones precisas para asegurar la seguridad de los trabajadores. Aunque la secuencialidad de los datos industriales se ha aprovechado en los últimos años para mejorar el rendimiento de los modelos de detección basados en Machine Learning y Deep Learning, la mayoría de métodos de explicabilidad ignoran este fenómeno, lo que puede conducir a explicaciones erróneas. Para abordar esta limitación, en este trabajo se propone Temporal Shapley (T-SHAP), un módulo de explicabilidad diseñado para integrarse en frameworks de detección de anomalías industriales. T-SHAP preserva las dependencias temporales agrupando las mediciones por instante antes de calcular la importancia, permitiendo así identificar con precisión el momento determinante de la anomalía. La eficacia de T-SHAP se valida en el conjunto de datos Tennessee Eastman Process, demostrando su potencial para facilitar la mitigación de anomalías por parte de los operadores en contraste con las soluciones tradicionales.**

***Index Terms*—Explainable Artificial Intelligence, Shapley values, Anomaly Detection, Time Series, Industrial Internet of Things**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

En la era de la Industria 5.0, la integración de dispositivos inteligentes y sistemas interconectados ha transformado los procesos productivos, colocando la seguridad del trabajador como elemento central [1]. En el núcleo de esta transformación están los sensores y los actuadores, que se enmarcan en sistemas de Industrial Internet of Things (IIoT) para recopilar datos y facilitar la toma de decisiones gracias a tecnologías como la Inteligencia Artificial (IA) [2] y el Big Data [3].

No obstante, esta creciente conectividad también ha expuesto a las instalaciones industriales a riesgos cibernéticos cada vez más sofisticados [4], [5]. Para contrarrestar estas amenazas, se ha producido un cambio de paradigma: de los sistemas de detección de intrusos (IDS) tradicionales basados en firmas hacia enfoques de detección de anomalías (AD), que intentan identificar comportamientos inusuales modelando y estableciendo lo que se considera actividad normal. Las técnicas de Machine Learning (ML) y Deep Learning (DL) han demostrado ser muy efectivas en el campo del AD, especialmente cuando aprovechan el comportamiento secuencial inherente de los datos de sensores y actuadores para realizar predicciones. Sin embargo, la opacidad de la mayoría de estos modelos, comúnmente denominados de *caja negra*, limita la capacidad de los operadores para mitigar las anomalías detectadas, dificultando su adopción en la vida real.

Para superar esta limitación, en los últimos años ha ganado peso el campo de la Inteligencia Artificial Explicable (xAI), cuyo objetivo principal es que el usuario final comprenda las decisiones tomadas por modelos de ML/DL. Entre los métodos más utilizados en este campo se encuentran aquellos basados en los valores de Shapley [6], que ofrecen un enfoque sistemático para cuantificar la contribución de cada característica de entrada en una predicción de un modelo. Sin embargo, a pesar de su popularidad, estos métodos suelen pasar por alto las dependencias temporales intrínsecas en los datos, lo que puede derivar en explicaciones incompletas o incluso engañosas. Abordar esta brecha resulta fundamental para garantizar que los frameworks de AD proporcionen alertas tanto precisas como comprensibles, mejorando la fiabilidad de las medidas de ciberseguridad industrial.

Para superar estas limitaciones, este trabajo introduce **Temporal Shapley (T-SHAP)**, un método de xAI basado en los valores Shapley diseñado para modelos de ML/DL que trabajen con series temporales y aplicable a frameworks de AD en IIoT. Las principales contribuciones de este trabajo son:

- **Módulo xAI Open Source**: Se presenta T-SHAP como un módulo de xAI diseñado para integrarse en frameworks de AD que operen con modelos de ML/DL sobre datos de series temporales. Este módulo es de código abierto y está disponible en [7].
- **Agrupación antes del cálculo**: Se introduce un novedoso enfoque de agrupación de características aplicado previo al cálculo de la importancia. Esta estrategia permite preservar las relaciones temporales, ignoradas por los métodos basados en los valores Shapley convencionales.
- **Validación en un escenario realista**: El método T-SHAP se evalúa sobre el conjunto de datos Tennessee Eastman Process (TEP) [8], un conjunto de datos industrial ampliamente utilizado en el campo del AD que simula un proceso químico complejo con diversos tipos de anomalías. Se compara el rendimiento de T-SHAP con el de SHAP, el módulo de xAI basado en los valores Shapley más utilizado; observando que T-SHAP, a diferencia de SHAP, es temporalmente consistente en sus explicaciones.

El resto del artículo está estructurado de la siguiente manera: La sección II trata la base teórica de los valores Shapley, la sección III revisa el estado del arte de los métodos

de xAI basados en los valores Shapley y la integración de módulos de xAI en frameworks de AD en IIoT, en la sección IV se presenta el módulo de xAI propuesto y se integra en un framework de AD, en la sección V se detallan los resultados experimentales y finalmente en la sección VI se discuten las conclusiones y el trabajo futuro.

II. Marco teórico

El método de los valores Shapley, originalmente propuesto en la teoría de juegos cooperativos, se ha adaptado al campo de la xAI para medir la contribución individual de cada característica de entrada sobre la predicción de un modelo de ML/DL. En un juego cooperativo, donde un conjunto de jugadores N colabora para alcanzar un resultado, se define una función $v : 2^N \to \mathbb{R}$ que asigna un valor a cada coalición $S \subset N$. El valor $\varphi_i(v)$ representa la contribución individual del jugador i y debe cumplir los siguientes axiomas [9]:

- **Simetría**: Si dos jugadores i y j contribuyen por igual a todas las coaliciones, son igual de importantes. Formalmente, si para todos los subconjuntos $S \subseteq N \setminus \{i, j\}$ se cumple que $v(S \cup \{i\}) = v(S \cup \{j\})$, entonces $\varphi_i(v) = \varphi_j(v)$.
- **Aditividad**: Si se consideran dos juegos independientes v y w, entonces para cada jugador i, se debe cumplir que $\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$.
- **Eficiencia**: La suma de las importancias de cada jugador debe ser igual al valor total del juego cuando todos los jugadores participan, $\sum_{i \in N} \varphi_i(v) = v(N)$.
- **Jugador ficticio**: Si un jugador i no altera el valor de cualquier coalición más allá de su propio valor individual, es decir, si para todo subconjunto $S \subseteq N \setminus \{i\}$, $v(S \cup i) = v(S) + v(i)$, entonces $\varphi_i(v) = v(i)$.

Lloyd Shapley demostró en [9] la existencia y unicidad de este valor, que además se puede calcular según la *Ec. (1)*.

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \cdot (v(S \cup \{i\}) - v(S)) \quad (1)$$

Esta ecuación evalúa para cada jugador i, su importancia sobre el resultado final. Para ello, estudia su contribución marginal a cada coalición S que inicialmente no lo contenga, ponderando esta según la probabilidad de formación de dicha coalición. En la adaptación de este método al campo de la xAI, se considera el vector de entrada x de un modelo y, para cualquier subconjunto de características S, se define x^S como el subvector de x que contiene los valores correspondientes a S y $x^{\overline{S}}$ el de los valores que no están en S. La contribución de S en una instancia específica x_* se cuantifica mediante la función de valor $v(S, x_*) = \mathbb{E}\big(f(x) \mid x^S = x_*^S\big)$, donde $f(\cdot)$ representa la función de predicción del modelo. Mediante este método se puede descomponer la predicción de un modelo de ML/DL en la contribución individual de cada una de las características que toma como entrada.

III. Estado del arte

Esta sección revisa el estado del arte en los métodos de xAI basados en los valores Shapley y los módulos de xAI integrados en frameworks de AD en IIoT.

III-A. Métodos de xAI basados en los valores Shapley

El uso de métodos de xAI basados en los valores Shapley en modelos de ML/DL enfrenta dos desafíos principales. Por un lado, la distribución del modelo a explicar es generalmente desconocida, lo que dificulta la estimación precisa de las características faltantes necesarias para calcular la función de valor de cada coalición. Por otro, el cálculo exacto de los valores Shapley implica una alta complejidad computacional, ya que requiere sumar hasta 2^{p-1} términos, siendo p el número de características.

Para la estimación de características faltantes se han propuesto diversas estrategias. Una opción consiste en utilizar un valor de referencia, como un vector de ceros o la media de las características, aunque esta elección puede introducir sesgos significativos [10], [11]. Alternativamente, se puede estimar empíricamente la distribución de probabilidad, lo que da lugar a dos enfoques: el condicionado y el marginal. En el método condicionado, el valor de cada coalición se calcula según el valor esperado del modelo $f(\cdot)$ condicionado a un conjunto específico de características S en un punto específico x_* según la fórmula $\mathbb{E}_{p(x_*^S | x_k^{\bar{S}})}[f(x_*^S, x_k^{\bar{S}})]$. Aquí,$x_k^{\bar{S}}$ representa las características que no están en S de la muestra k-ésima del conjunto soporte, definido como un pequeño grupo de muestras representativas del conjunto de datos. Este método puede implementarse asumiendo distribuciones paramétricas, como la gaussiana [12], utilizando modelos generativos o surrogados [13], o incluso entrenando el modelo para que acepte directamente la coalición como entrada del modelo de AD [14]. En contraste, el enfoque marginal calcula el valor esperado del modelo sobre la distribución marginal de S, lo que equivale a emplear el valor medio de cada característica como valor de referencia y se expresa mediante $\mathbb{E}_{p(x^S)}[f(x_*^S, x_k^{\bar{S}})]$ [15].

Respecto a la complejidad computacional, diversas técnicas de aproximación han sido desarrolladas para evitar el cálculo exhaustivo de todos los subconjuntos. Entre estas se encuentran métodos de muestreo que consideran solo ciertos subconjuntos de características, como [16], en donde cada subconjunto S se selecciona con una probabilidad $P(S) = \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!}$. De manera alternativa, en [6] se transforma el problema en uno de minimización de mínimos cuadrados, aplicando técnicas de regularización para reducir el número de características.

Finalmente, se han desarrollado métodos específicos para ciertos tipos de modelos. Por ejemplo, TreeSHAP aprovecha la estructura de modelos basados en árboles para evaluar rápidamente las contribuciones de las características [17], mientras que en modelos lineales, bajo la asunción de independencia entre características, es posible obtener soluciones cerradas mediante el método LinearSHAP [6].

III-B. Módulos de xAI en frameworks de AD

La AD en IIoT ha demostrado ser un campo prometedor, ya que los modelos de ML/DL son capaces de aprovechar la información temporal de los datos provenientes de sensores y actuadores para alcanzar un buen rendimiento [18]. Sin embargo, los framework de AD actuales presentan dos limitaciones: la ausencia de módulos de xAI o la implementación

de métodos de xAI que no consideran la secuencialidad de los datos, lo que dificulta su adopción en entornos críticos.

Muchas de los frameworks de AD que han sido propuestos, a pesar de obtener buenos resultados, comparten la carencia de técnicas de métodos xAI. Por ejemplo, en [19] se introduce una arquitectura basada en un modelo LSTM (Long Short-Term Memory) y CNN (Convolutional Neural Network) para reconstruir las características de entrada, validada con el conjunto de datos UNSW-NB15 [20]. De manera similar, se han estudiado arquitecturas combinadas con redes de tipo MLP (MultiLayer Perceptron) [21] y modelos de AutoEnconders (AE) [22]. Aunque todos estos estudios presentan un buen rendimiento en AD, no ofrecen ningún método para comprender las causas de las anomalías, dificultando su adopción en entornos reales.

Ante esta limitación, han surgido técnicas de xAI que dotan de explicabilidad a los frameworks de AD. Estas técnicas pueden clasificarse según su objetivo en técnicas globales, las cuales buscan explicar el comportamiento general del modelo y técnicas locales, que se centran en explicar predicciones individuales. En el primer caso, el método PDP (Partial Dependence Plot) muestra la relación promedio entre una o dos características y la respuesta del modelo, mientras que el método ALE (Accumulated Local Effects) también describe cómo una variable influye en la predicción promedio, pero lo hace dividiendo el rango de esa variable en intervalos y midiendo localmente cómo cambian las predicciones. En el caso de las técnicas locales, se encuentran tanto los métodos basados en los valores Shapley ya explicados anteriormente, donde destaca SHAP [6], como el método LIME (Local Interpretable Model-Agnostic Explanations). LIME aproxima localmente el comportamiento del modelo mediante modelos más simples que permiten evaluar más fácilmente la importancia de cada característica en la predicción.

Estas técnicas de xAI se han comenzado a integrar recientemente en frameworks de AD sobre entornos IIoT. Por ejemplo, en [23] se utiliza SHAP para explicar un modelo híbrido CNN-LSTM que predice la vida útil restante de motores, calculando la contribución de cada medición de cada sensor en ventanas temporales y sumando estos valores para identificar el sensor responsable de una anomalía. De forma similar, en [24] se combina un enfoque de AD basado en ECOD (Empirical Cumulative-distribution-based Outlier Detection) y DeepSVDD (Deep Support Vector Data Description) con técnicas de xAI como SHAP, LIME y ALE, acompañado de un procedimiento de agrupamiento a posteriori de las importancias de las características. Además, estudios como los de [25] y [26] integran SHAP en frameworks basados en modelos de AD de tipo LSTM (a través de un AE-LSTM y una Bidirectional-LSTM, respectivamente) calculando el valor Shapley de cada característica para después asignar una contribución individual a cada sensor. No obstante, la agregación a posteriori de las importancias puede no capturar adecuadamente las dependencias temporales inherentes, lo que evidencia la necesidad de métodos xAI específicos para datos secuenciales.

IV. Integración de T-SHAP en un framework de AD

En esta sección se introduce el módulo xAI propuesto, Temporal SHAP (T-SHAP), diseñado para mejorar la explicabilidad en frameworks de AD que emplean modelos ML/DL con datos dispuestos en forma de series temporales. La sección se divide en dos subsecciones: la primera describe un framework típico de AD y la segunda la integración de T-SHAP en este.

IV-A. Definición de un framework de AD

El framework de AD es el encargado de recibir los datos de sensores y actuadores del sistema IIoT, transformarlos en un formato consistente y crear un modelo de AD para detectar comportamientos anómalos en estos. Un framework de este tipo se divide en dos módulos, los cuales se explicarán a continuación.

IV-A1. Módulo de preparación de los datos: Este módulo se encarga de transformar los datos al formato que espera el detector de anomalías. En primer lugar, se exploran y limpian los datos, corrigiendo el ruido en las mediciones y ajustando aquellos casos mal etiquetados o que no reflejan anomalías reales. Posteriormente, se dividen los datos en tres conjuntos de datos: entrenamiento, validación y test. El conjunto de entrenamiento se utilizará para que el modelo aprenda los patrones subyacentes de los datos, el de validación servirá para ajustar los hiperparámetros y prevenir el sobreajuste y el de test se reserva para evaluar la capacidad de generalización del modelo en datos no vistos. Tras esta partición, se escalan las variables continuas a un mismo rango para evitar que alguna domine el proceso de aprendizaje. Finalmente, para aprovechar la secuencialidad de los datos, se organizan en ventanas temporales, es decir, en conjuntos de mediciones consecutivas. El parámetro *stride* determina el desplazamiento entre una ventana y la siguiente (*Fig 1*).

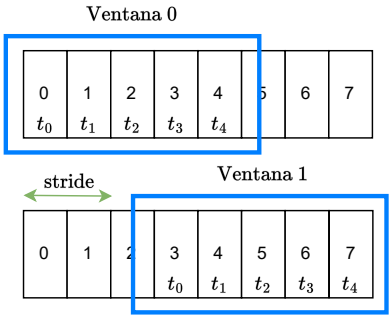


Figura 1. Creación de dos ventanas temporales con $stride = 2$

IV-A2. Módulo de generación del modelo de AD: La elección del modelo de AD depende de las características de los datos y de la disponibilidad de etiquetas: si las hay, se recurre a métodos supervisados para distinguir entre estados normales y anómalos; en su ausencia, se utilizan modelos no supervisados, como AEs. Además, el modelo seleccionado debe manejar datos secuenciales, por lo que se opta habitualmente por redes LSTM o CNN. Una vez elegido el modelo de AD, se ajustan los hiperparámetros que definen tanto su estructura (por ejemplo, número de capas o neuronas)

como el proceso de entrenamiento (por ejemplo, tasa de aprendizaje). Para ello, se especifican varios valores posibles para cada hiperparámetro y se entrena sobre el conjunto de entrenamiento un modelo de AD con cada combinación, evaluando el rendimiento en el conjunto de validación para seleccionar la mejor configuración. Dado que el espacio de búsqueda suele ser amplio, se suele utilizar un método *coarse-to-fine*: primero se explora una parrilla amplia de valores para identificar una región prometedora y, posteriormente, se define una parrilla más fina en torno a dicha región para encontrar el óptimo. Durante el entrenamiento se minimiza una función de pérdida seleccionada. En este tipo de frameworks, es habitual utilizar la *cross entropy loss*, que cuantifica la discrepancia entre la distribución real y la predicha. En problemas de clasificación, se define según la *Ec. (2)* donde y_i es la etiqueta real y $\hat{y}_i$ la probabilidad de pertenencia a esta clase asignada por el modelo. En escenarios multiclase, la función se adapta sumando sobre todas las clases, permitiendo evaluar de forma simultánea la probabilidad de cada categoría.

$$Error = -\sum_{i \in Entrenamiento} y_i \log(\hat{y}_i) \tag{2}$$

Finalmente, el modelo se evalúa en el conjunto de test. Para ello, se definen los siguientes conceptos: los verdaderos positivos (TP) son las anomalías detectadas correctamente, los falsos positivos (FP) las normales clasificadas como anómalas, los falsos negativos (FN) las anomalías no detectadas y los verdaderos negativos (TN) las normales reconocidas adecuadamente. Con estos valores se calculan las métricas indicadas en la *Ec. (3)*, donde Precision indica qué proporción de detecciones son realmente anomalías, Recall mide cuántas anomalías reales se detectan y F1-score es la media armónica de ambas. En problemas multiclase, estas métricas se determinan para cada clase y se promedian para obtener una valoración global del rendimiento.

$$\begin{aligned} \text{Precisión} &= \frac{TP}{TP+FP}, \\ \text{Recall} &= \frac{TP}{TP+FN}, \\ \text{F1-score} &= \frac{2 \cdot \text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}. \end{aligned} \tag{3}$$

IV-B. Propuesta de módulo de xAI: T-SHAP

Una vez definido el framework de AD, se presenta la contribución principal de este trabajo: el módulo de xAI **Temporal SHAP (T-SHAP)**. Este método de código abierto disponible en [7], ha sido diseñado específicamente para integrarse en frameworks de AD aplicados a IIoT. La principal novedad de T-SHAP radica en su capacidad para capturar la secuencialidad de los datos y proporcionar explicaciones contextualizadas. Para ello, el módulo agrupa las características de cada ventana que toma como entrada el modelo de AD, creando un grupo por cada instante de esta. Esta agrupación permite evaluar la influencia conjunta de las mediciones de sensores y actuadores en cada instante mediante el cálculo de su valor Shapley como grupo. Este enfoque, a diferencia de la agrupación a posteriori, la cual consiste en calcular el valor Shapley para cada sensor o actuador en cada instante para sumarlos y obtener la importancia de este grupo, preserva de forma nativa la dinámica temporal de los datos. Gracias a la integración de T-SHAP, los operadores pueden identificar con precisión el instante determinante de una alerta y tomar las medidas de mitigación adecuadas.

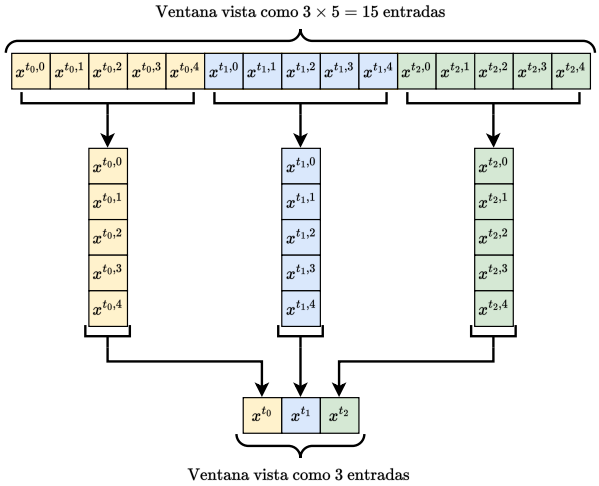


Figura 2. Agrupación de las características de una ventana según instante.

El módulo se divide en tres componentes, los cuales se explicarán a continuación.

IV-B1. Componente de agrupación de las características: Este componente es el encargado de agrupar las características de entrada del modelo de AD (*Fig 2*). Se denotará como x una ventana de tiempo, formada por p mediciones de sensores/actuadores a lo largo de w instantes de tiempo. Para hacer referencia al valor de la medición j en un instante i se denotará $x^{t_i,j}$. Para denotar el vector de todas las mediciones de un instante i, se usará x^{t_i} y para referirnos a solo un subconjunto de estas $S \subseteq \{0, \ldots, p-1\}$, se escribirá $x^{t_i,S}$. El espacio de características, de tamaño $w \times p$, se denota $N = \{(t_0,0),...(t_0,p-1),...,(t_{w-1},0),...,(t_{w-1},p-1)\}$, cuyo número de subconjuntos total es $|\mathcal{P}(N)| = 2^{w \times p}$. Mediante esta agrupación, se pasa de ver la entrada del modelo como $x = (x^{t_0,0}, \ldots, x^{t_0,p-1}, \ldots, x^{t_{w-1},p-1})$ a hacerlo como $x = (x^{t_0}, \ldots, x^{t_{w-1}})$. Así, el nuevo espacio de características $N_w = \{t_0, ..., t_{w-1}\}$ tiene un tamaño w y el número de subconjuntos total es $|\mathcal{P}(N_w)| = 2^w$.

IV-B2. Componente de cálculo de la función valor: Este componente será el encargado de determinar para cada ventana x_* y cada coalición S de instantes de esta una función de valor $v(x_*, S)$. Esta función se define teóricamente como el valor esperado del modelo condicionado a los instantes de tiempo S en la ventana x_* (denotado como x_*^S), esto es, $v(S, x_*) = \mathbb{E}(f(x)|x^S = x_*^S)$. En la práctica, este cálculo es inviable porque implica conocer la distribución de probabilidad del modelo $f(\cdot)$. Para resolver esto, se decide estimar directamente la función creando un conjunto representativo de las ventanas utilizadas para entrenar el modelo de cierto tamaño K, también llamado conjunto soporte. A la hora de calcular el valor de x_*^S, se crean K ventanas formadas por la combinación de los instantes S de x_* y los $\bar{S}$ de cada una de las ventanas del conjunto soporte. Tras esto, se evalúa el

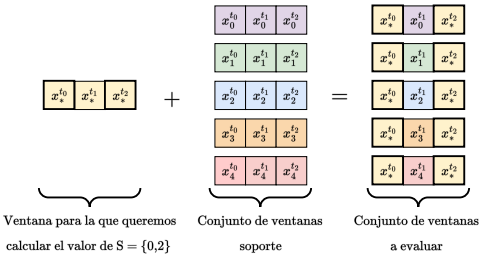


Figura 3. Uso del conjunto soporte para la estimación del valor de una coalición

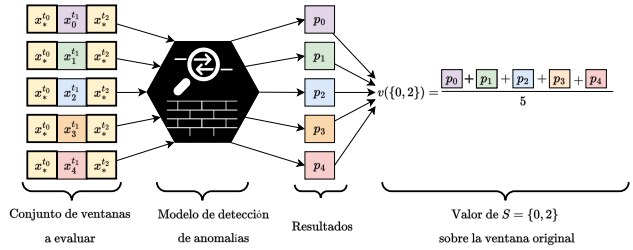


Figura 4. Estimación del valor de una coalición a partir de las ventanas creadas

modelo de AD en estas ventanas, siendo la función valor el promedio de las probabilidades de anomalía obtenidas según la *Ec. (4)*. (*Fig. 3* y *Fig. 4*)

$$v(S, x_*) \approx \frac{1}{K} \sum_{k=1}^{K} f(x_k^{\overline{S}}, x_*^S) \tag{4}$$

IV-B3. Componente de cálculo de los valores T-SHAP: Una vez que las características han sido agrupadas según instante y se ha decidido el criterio para la estimación de la función valor, se procede a calcular el valor Shapley de cada grupo. En la *Ec. (5)* se muestra la fórmula para el cálculo del valor Shapley de cada grupo. Debido a la agrupación realizada, esta tiene una complejidad de 2^w, frente a la complejidad de $2^{w \cdot p}$ que tendría la fórmula original sin agrupaciones. Pese a que 2^w es una complejidad viable para ventanas pequeñas, puede resultar prohibitivo en otros casos. Para abordar este problema se presenta además un método aproximado en la *Ec. (6)* que consiste en muestrear M_w subconjuntos de $N_w \setminus \{i\}$, donde cada subconjunto S_w se selecciona con una probabilidad $P(S) = \frac{|S|! \cdot (|N|-|S|-1)!}{|N|!}$ según el método presentado en [16].

$$\varphi_i(v, x_*) = \sum_{S_w \subseteq N_w \setminus \{i\}} \frac{|S_w|! \cdot (|N_w| - |S_w| - 1)!}{|N|!} \cdot \Big(v(S_w \cup \{i\}, x_*) - v(S_w, x_*) \Big) \tag{5}$$

$$\varphi_i(v, x_*) = \sum_{S_w \in M_w} \left(v(S_w \cup \{i\}, x_*) - v(S_w, x_*) \right) \tag{6}$$

V. Resultados experimentales

Esta sección detalla los experimentos llevados a cabo para comprobar el funcionamiento de T-SHAP. Para ello, se definirá un framework de AD sobre un conjunto de datos industrial, se integrará T-SHAP en este y finalmente se validará

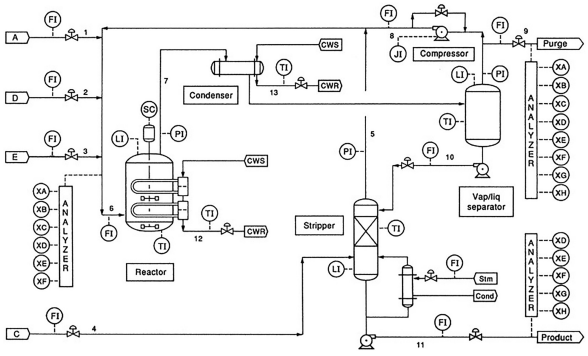


Figura 5. Esquema del proceso químico simulado por TEP

su capacidad de tener en cuenta los patrones secuenciales de los datos comparando su funcionamiento con el de SHAP, la librería basada en los valores Shapley más utilizada dentro del campo de la xAI [6].

V-A. Creación de un framework de AD

A continuación se explica el proceso de creación del framework de AD.

V-A1. Conjunto de datos utilizado: El conjunto de datos empleado en la experimentación es TEP [8], [27], ampliamente utilizado en la AD. Este conjunto proviene de la simulación de un proceso químico (*Fig. 5*), en el que se introducen 20 tipos de fallos que se consideran anomalías. El sistema es monitorizado por 52 sensores/actuadores, que realizan mediciones cada 3 minutos. TEP consta de 1 000 simulaciones para cada una de sus 21 clases (una clase normal y 20 anómalas), estructuradas en 500 simulaciones para entrenamiento y 500 para test. Cada simulación normal consta de 500 mediciones, mientras que las simulaciones con anomalías contienen 960 mediciones.

V-A2. Módulo de preparación de datos: Para el tratamiento de datos se adoptó la siguiente metodología. En primer lugar, se excluyeron las anomalías 3, 9 y 15, ya que la perturbación en estas no es suficiente como para generar una anomalía [28]. De cada una de las 500 simulaciones por clase destinadas al entrenamiento, se utilizaron 400 para el conjunto de entrenamiento y 100 para el de validación. El conjunto de test se compone de la unión de las 500 simulaciones restantes para cada clase. En las simulaciones relativas a clases anómalas se eliminaron las primeras 160 muestras, ya que representan tráfico normal previo a la introducción de la anomalía. Las mediciones de los 52 sensores/actuadores son representadas mediante características continuas. Se calcularon la media y la desviación estándar de cada característica en el conjunto de entrenamiento, y estos parámetros se emplearon para escalar de forma homogénea los conjuntos de entrenamiento, validación y test. Finalmente, se dividieron los conjuntos de datos en ventanas temporales de tamaño 10 con un stride de 5. Esto implica que, en cada instante, el modelo recibe 52×10 características de entrada, correspondientes a las últimas 10 mediciones (equivalentes a 30 minutos) de cada sensor/actuador.

V-A3. Módulo de AD: El modelo de AD elegido fue una red neuronal con neuronas de tipo LSTM, capaces de capturar patrones secuenciales a corto y largo plazo. La función de

error utilizada fue la cross entropy adaptada a un problema de clasificación multiclase. Los hiperparámetros a optimizar en este modelo incluyen el tamaño oculto, que determina la cantidad de neuronas en cada capa; el número de capas, que define la profundidad de la red; la dirección bidireccional, que permite procesar la secuencia en ambas direcciones; y el dropout, que actúa como regularizador para prevenir el sobreajuste. Se llevó a cabo una búsqueda de hiperparámetros mediante un enfoque coarse-to-fine, iniciando con una exploración sobre la parrilla amplia de hiperparámetros de la *Tabla I* para refinar después los rangos de valores hasta alcanzar la configuración óptima encontrada sobre la parrilla fina de la *Tabla II*. La evaluación del modelo en el conjunto de test mostró un buen rendimiento, con promedios macro de 0.98 en las métricas de F1-Score, Recall y Precision, como se puede ver en la *Tabla III*.

Tabla I
PARRILLA DE HIPERPARÁMETROS AMPLIA. LOS VALORES ÓPTIMOS ESTÁN EN NEGRITA

Hiperparámetros	Valores
Tamaño oculto	[8, **32**, 128]
Número de capas	[1, **5**, 9]
Bidireccional	[**True**, False]
Dropout	[**0**, 0.1, 0.2]

Tabla II
PARRILLA DE HIPERPARÁMETROS FINA. LOS VALORES ÓPTIMOS ESTÁN EN NEGRITA

Hiperparámetros	Valores
Tamaño oculto	[16, **32**, 64]
Número de capas	[**3**, 5, 7]
Bidireccional	[**True**]
Dropout	[**0**]

V-B. Configuración del módulo T-SHAP

Una vez se entrenó el modelo de AD, el siguiente paso era integrar el módulo T-SHAP en este. En primer lugar, el componente de agrupación divide cada ventana de entrada en 10 grupos, uno por instante. De este modo, se puede evaluar la influencia sobre la predicción del modelo de AD de cada instante de una ventana. En segundo lugar, se construye el conjunto soporte para la estimación de la función valor. Para ello, se seleccionaron 10 ventanas del conjunto de entrenamiento para cada clase; considerando 18 clases, el conjunto soporte consta de 180 instancias. Esta estratificación del conjunto soporte garantiza que cada clase esté representada de manera equilibrada, lo que asegura una correcta estimación. Finalmente, dado que el tamaño de la ventana es 10, el número total de subconjuntos a evaluar para cada ventana es $2^{10} = 1024$, lo cual no resulta computacionalmente prohibitivo. Por ello, se optó por utilizar el método exacto para el cálculo de los valores T-SHAP.

Tabla III
MÉTRICAS DEL MODELO DE AD SOBRE EL CONJUNTO DE TEST

Clase	Precision	Recall	F1-Score
0	0.94	0.86	0.90
1	1	1	1
2	1	1	1
4	1	1	1
5	0.99	1	0.99
6	1	0.99	1
7	1	1	1
8	0.99	0.99	0.99
10	0.97	0.97	0.97
11	0.98	0.99	0.99
12	0.99	0.89	0.94
13	0.99	0.97	0.98
14	1	1	1
16	0.99	0.96	0.98
17	1	1	1
18	0.90	0.99	0.95
19	0.99	0.97	0.98
20	0.88	1	0.93
Accuracy		0.98	
Macro avg	0.98	0.98	0.98

V-C. Ratificación de la consistencia temporal de T-SHAP

Una vez configurado el framework de AD e integrado el módulo de xAI T-SHAP, se realizaron experimentos para evaluar su capacidad de preservar las dependencias temporales de los datos, en contraste con otros enfoques existentes. Con este fin, se usó SHAP, el método de xAI basado en los valores Shapley de referencia [6]. Dentro de la librería de SHAP hay distintos métodos de aproximación pero se decidió utilizar KernelSHAP por ser el más mencionado en la literatura. Para este módulo, se configuró el mismo conjunto soporte creado para T-SHAP y KernelSHAP. A diferencia de T-SHAP, que agrupa las mediciones por instante antes de calcular los valores Shapley, para hacer esto mediante KernelSHAP se deben calcular los valores Shapley de las mediciones de cada sensor/actuador en cada instante para después sumar a posteriori los correspondientes a un mismo instante para determinar su importancia.

Para comprobar la capacidad de ambos métodos de preservar las dependencias temporales de los datos se analizó la consistencia temporal de sus explicaciones. En base a esto, se definió la siguiente hipótesis: si dos ventanas consecutivas difieren únicamente en un instante, las importancias asignadas a los instantes comunes deberían mantenerse parecidas aunque su posición dentro de la ventana varíe. Así, si en la primera ventana el instante t_2 es el más influyente, en la ventana siguiente el instante más influyente debería ser t_1, a menos que el nuevo instante t_9 influya sustancialmente sobre la predicción del modelo. Para poner a prueba esta hipótesis, se creó un conjunto de datos con *stride* de 1 (las ventanas consecutivas se desplazan tan solo un instante) a partir del conjunto de test, incluyendo una simulación, elegida de manera aleatoria para cada una de las 18 clases (1 normal y 17 anómalas). En las visualizaciones diseñadas para este experimento, el eje x representa ventanas consecutivas, el eje y izquierdo los instantes temporales de la ventana y el eje y derecho muestra la probabilidad de pertenencia a la clase predicha por el modelo de AD. Cada celda de la gráfica indica la contribución de un instante dentro de una ventana, con tonalidades rojas para influencias positivas (más intensas cuanto mayor es la influencia) y azules para influencias negativas, mientras que las celdas sin color denotan escasa relevancia. La hipótesis de validación se puede consultar en la *Fig. 6*.

Tal como se detalla en la Tabla IV, T-SHAP exhibe una consistencia temporal clara en 10 de las 18 clases analizadas,

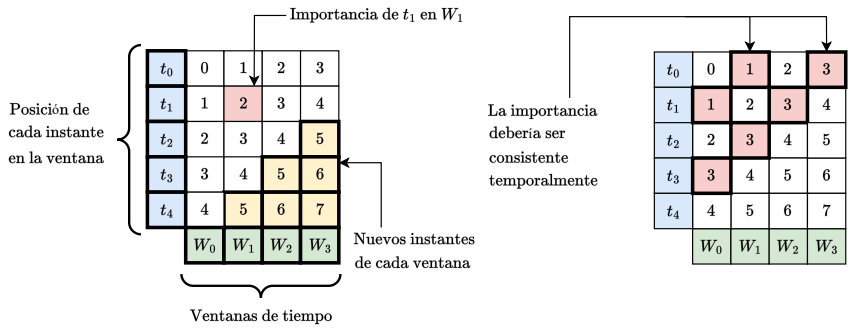


Figura 6. Método de ratificación de la consistencia temporal definido, dependiente de la aparición de patrones diagonales en los gráficos generados

Tabla IV
CLASIFICACIÓN DE LAS CLASES DE TEP EN FUNCIÓN DE LA CONSISTENCIA TEMPORAL OBSERVADA PARA CADA UNA POR PARTE DE T-SHAP Y SHAP

Consistencia temporal	SHAP	T-SHAP
Clara	∅	1, 2, 4, 5, 7, 11 12, 14, 19, 20
Difusa	∅	8, 10, 13, 17
No apreciable	0, 1, 2, 4, 5, 6 7, 8, 10, 11, 12 13, 14, 16, 17, 18 19, 20	0, 6, 16, 18

difusa en 4 clases y no apreciable en otras 4. En cambio, SHAP con agrupación a posteriori no mostró patrones temporales consistentes, presentando un comportamiento ruidoso e inconsistente entre ventanas.

Las *Fig. 7* y *Fig. 8* ilustran el comportamiento de SHAP y T-SHAP en las primeras 60 ventanas de las simulaciones correspondientes a las anomalías 19 y 17 respectivamente. En la parte inferior de la *Fig. 7* (T-SHAP) se observan patrones diagonales nítidos, reflejo de una consistencia temporal clara; por el contrario, en la parte superior (SHAP) predomina el ruido y los cambios bruscos en las atribuciones, lo que impide identificar los instantes más influyentes. En la *Fig. 8*, que representa la simulación relativa a la anomalía 17, se aprecia un ejemplo de consistencia temporal más difusa en la parte inferior (T-SHAP). Además aquí, entre las ventanas 20 y 30, los instantes recién incorporados (situados al final de la ventana) apenas influyen en la predicción, mientras que los instantes más antiguos, con colores rojos, ejercen mayor impacto. Una vez que estos instantes dejan de formar parte de la ventana, la probabilidad de anomalía desciende de 0.99 a 0.7, evidenciando la importancia que estos tenían sobre el resultado del modelo.

Estos resultados confirman la capacidad de T-SHAP de interpretar de manera correcta modelos de AD que trabajan con datos de sistemas IIoT dispuestos en forma de series temporales, frente a la incapacidad de los métodos tradicionales como SHAP.

VI. CONCLUSIONES Y VÍAS FUTURAS

En este trabajo se evidenció que los métodos actuales de xAI basados en los valores Shapley no resultan adecuados para modelos de AD sobre IIoT donde los datos de entrada se estructuran en ventanas temporales. Para proporcionar explicaciones a los operadores para mitigar las anomalías detectadas, estos métodos suelen calcular el valor Shapley de cada característica de la ventana de tiempo y agrupar a posteriori estas importancias, ignorando así las dependencias temporales inherentes a estos datos. Para solventar este problema, se propuso T-SHAP, un módulo de xAI que realiza la agrupación por instante antes de calcular los valores Shapley, preservando así las dependencias temporales y permitiendo determinar de forma precisa el instante detonante de una alerta. Para probar su eficacia, se definió un framework de AD sobre el conjunto de datos TEP y se integró T-SHAP sobre este. Una vez hecho esto, se compararon los resultados proporcionados por T-SHAP con los de SHAP, evidenciando la capacidad de T-SHAP de preservar dependencias temporales frente a SHAP.

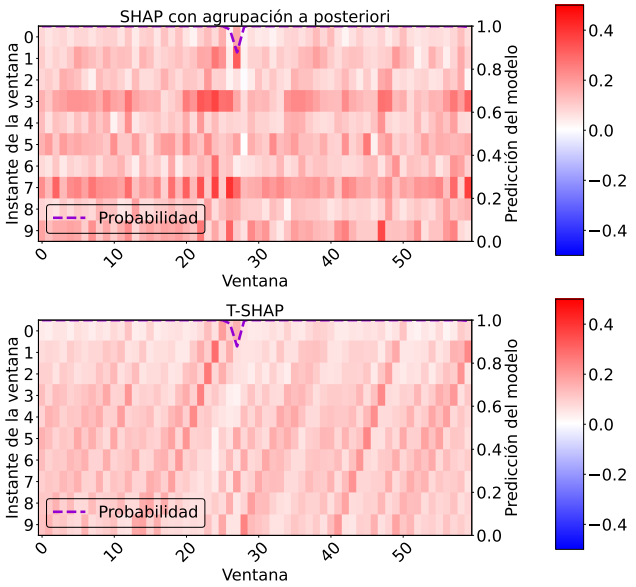


Figura 7. Valores SHAP con agrupación a posteriori (arriba) y T-SHAP (abajo) para las primeras 60 ventanas de la simulación de la anomalía 19.

Como líneas futuras, se plantea explorar nuevas modalidades de agrupación (por sensor/actuador o proceso industrial) que proporcionen información más detallada para los operadores, así como implementar métodos adicionales para el cálculo de T-SHAP que reduzcan su complejidad computacional sin menoscabar su capacidad explicativa.

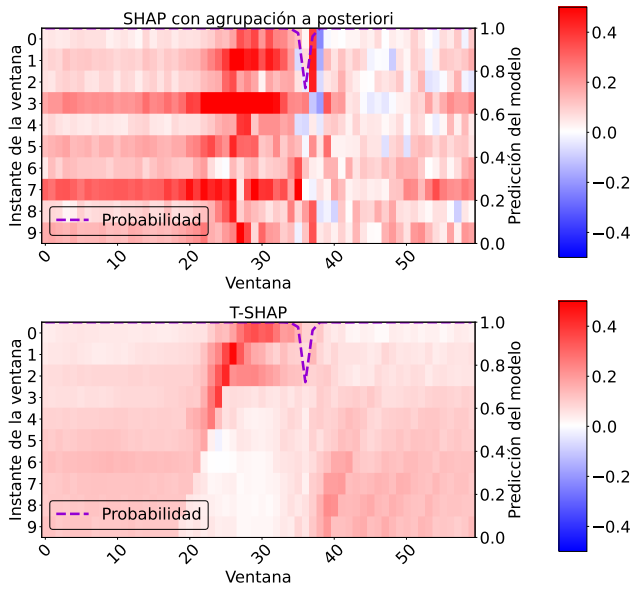


Figura 8. Valores SHAP con agrupación a posteriori (arriba) y T-SHAP (abajo) para las primeras 60 ventanas de la simulación de la anomalía 17

VII. Agradecimientos

Este trabajo ha sido financiado por (a) Grant TED2021-129300B-I00, mediante MCIN/AEI/10.13039/501100011033, NextGeneration EU/PRTR, UE, (b) Grant PID2021-122466OB-I00, mediante MCIN/AEI/10.13039/501100011033/FEDER, UE, y (c) el proyecto estratégico DEFENDER del Instituto Nacional de Ciberseguridad de España (INCIBE), mediante el plan Next Generation EU de Recuperación, Transformación y Resiliencia.

Referencias

[1] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and industry 5.0—inception, conception and perception," *Journal of manufacturing systems*, vol. 61, pp. 530–535, 2021.

[2] J. Lee, H. Davari, J. Singh, and V. Pandhare, "Industrial artificial intelligence for industry 4.0-based manufacturing systems," *Manufacturing letters*, vol. 18, pp. 20–23, 2018.

[3] L. D. Xu and L. Duan, "Big data for cyber physical systems in industry 4.0: a survey," *Enterprise Information Systems*, vol. 13, no. 2, pp. 148–169, 2019.

[4] U. P. D. Ani, H. He, and A. Tiwari, "Review of cybersecurity issues in industrial critical infrastructure: manufacturing in perspective," *Journal of Cyber Security Technology*, vol. 1, no. 1, pp. 32–74, 2017.

[5] K. E. Hemsley, E. Fisher *et al.*, "History of industrial control system cyber incidents," Idaho National Lab.(INL), Idaho Falls, ID (United States), Tech. Rep., 2018.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1705.07874

[7] M. F. de la Peña, "T-shap: A shapley-based xai module for time series ad models," 2024, gitHub repository. Available at: https://github.com/ManuelFranco/TSG-SHAP/tree/92bc5f33044952be5391e07d778e025ab2b0f1d9.

[8] C. A. Rieth, B. D. Amsel, R. Tran, and M. B. Cook, "Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation," 2017. [Online]. Available: https://doi.org/10.7910/DVN/6C3JR1

[9] L. Shapley, *A Value for n-Person Games*. Princeton: Princeton University Press, 1953, pp. 307–317. [Online]. Available: https://doi.org/10.1515/9781400881970-018

[10] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3145–3153. [Online]. Available: https://proceedings.mlr.press/v70/shrikumar17a.html

[11] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3449–3457. [Online]. Available: https://doi.org/10.48550/arXiv.1704.03296

[12] H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee, "True to the model or true to the data?" 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.16234

[13] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2006.01272

[14] I. Covert, S. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2011.14878

[15] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 9269–9278. [Online]. Available: https://proceedings.mlr.press/v119/sundararajan20b.html

[16] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Computers and Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009. [Online]. Available: https://doi.org/10.1016/j.cor.2008.04.004

[17] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[18] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882

[19] X. Zhou, Y. Hu, W. Liang, J. Ma, and Q. Jin, "Variational lstm enhanced anomaly detection for industrial big data," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 3469–3477, 2021. [Online]. Available: https://doi.org/10.1109/TII.2020.3022432

[20] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/MilCIS.2015.7348942

[21] Z. E. Huma, S. Latif, J. Ahmad, Z. Idrees, A. Ibrar, Z. Zou, F. Alqahtani, and F. Baothman, "A hybrid deep random neural network for cyberattack detection in the industrial internet of things," *IEEE Access*, vol. 9, pp. 55 595–55 605, 2021. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3071766

[22] I. A. Khan, M. Keshk, D. Pi, N. Khan, Y. Hussain, and H. Soliman, "Enhancing iiot networks protection: A robust security model for attack detection in internet industrial control systems," *Ad Hoc Networks*, vol. 134, p. 102930, 2022. [Online]. Available: https://doi.org/10.1016/j.adhoc.2022.102930

[23] C. W. Hong, C. Lee, K. Lee, M.-S. Ko, and K. Hur, "Explainable artificial intelligence for the remaining useful life prognosis of the turbofan engines," in *2020 3rd ieee international conference on knowledge innovation and invention (ickii)*. IEEE, 2020, pp. 144–147.

[24] K. Mathuros, S. Venugopalan, and S. Adepu, "Waxai: Explainable anomaly detection in industrial control systems and water systems," in *Proceedings of the 10th ACM Cyber-Physical System Security Workshop*, 2024, pp. 3–15.

[25] N. X. Hoang, N. V. Hoang, N. H. Du, T. T. Huong, K. P. Tran *et al.*, "Explainable anomaly detection for industrial control system cybersecurity," *IFAC-PapersOnLine*, vol. 55, no. 10, pp. 1183–1188, 2022.

[26] C. Hwang and T. Lee, "E-sfd: Explainable sensor fault detection in the ics anomaly detection system," *IEEE Access*, vol. 9, pp. 140 470–140 486, 2021.

[27] J. Downs and E. Vogel, "A plant-wide industrial process control problem," *Computers and Chemical Engineering*, vol. 17, no. 3, pp. 245–255, 1993. [Online]. Available: https://doi.org/10.1016/0098-1354(93)80018-I

[28] M. Onel, C. A. Kieslich, and E. N. Pistikopoulos, "A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the tennessee eastman process," *AIChE Journal*, vol. 65, no. 3, pp. 992–1005, 2019. [Online]. Available: https://doi.org/10.1002/aic.16497

Aprendizaje Federado Descentralizado y Conciencia Situacional para el Reconocimiento Aéreo Militar Seguro y Resiliente

Enrique Tomás Martínez Beltrán[1], Miguel Fernández Llamas[1], Anas Zine Boujemaoui[1], Gérôme Bovet[2], Gregorio Martínez Pérez[1], Alberto Huertas Celdrán[1]
[1]Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, 30100, España
Email: {enriquetomas, m.fernandezllamas, anas.z.b, gregorio, alberto.huertas}@um.es
[2]Cyber-Defence Campus, Armasuisse Science and Technology, 3602 Thun, Switzerland
Email: gerome.bovet@armasuisse.ch

***Resumen*—La guerra multidominio y en mosaico son estrategias militares que se basan en la integración de capacidades para lograr una ventaja estratégica, donde las misiones de reconocimiento aéreo juegan un papel crucial en la obtención de inteligencia operativa. En este contexto, las técnicas de aprendizaje automático han demostrado ser fundamentales para la detección precisa de vehículos militares. No obstante, estas misiones enfrentan desafíos significativos debido a los ataques adversarios dirigidos contra los modelos de aprendizaje y los vehículos aéreos, lo que compromete la integridad del reconocimiento y la seguridad de las operaciones. La combinación del Aprendizaje Federado Descentralizado (DFL) y la conciencia situacional podría mitigar estos desafíos, permitiendo el uso de modelos colaborativos, robustos y que preserven la privacidad. Sin embargo, la literatura existente pasa por alto la integración de estos enfoques y su aplicabilidad en misiones de reconocimiento militar. Este trabajo presenta NEBULA-SA, una extensión de la plataforma NEBULA que integra DFL con un módulo de defensa basado en la conciencia situacional para fortalecer la resiliencia ante ataques. La solución considera métricas clave como similitud de modelos, formaciones de vuelo, geoposicionamiento, uso de recursos y conectividad entre drones. Para evaluar su rendimiento, se diseñó un escenario militar simulado basado en datos de radar de apertura sintética (SAR), donde cuatro drones operan en formaciones tácticas bajo ataques dirigidos que afectaban su geoposicionamiento y colaboración. Los experimentos demostraron que NEBULA-SA supera enfoques previos al mejorar la precisión y la velocidad de detección sin comprometer la privacidad ni la resiliencia operativa.**

***Index Terms*—Aprendizaje Federado, Modelos Colaborativos, Descentralización, Conciencia Situacional, Mitigación, Seguridad**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Las operaciones militares modernas, caracterizadas por estrategias como la guerra en mosaico, requieren capacidades de reconocimiento rápidas, adaptables y precisas para mantener el dominio del campo de batalla [4]. Entre estas capacidades, el reconocimiento aéreo es fundamental para la detección de vehículos enemigos y la monitorización de sus actividades, lo que exige sistemas con inteligencia, robustez y adaptabilidad. El objetivo principal de estos sistemas es distinguir con precisión los objetivos de interés, minimizando la incertidumbre y garantizando la fiabilidad en la identificación. En este contexto, la integración de técnicas de ML en los sistemas de reconocimiento se ha convertido en un área clave de investigación, mejorando la rapidez, precisión y automatización en la detección de amenazas [5]. Sin embargo, estos sistemas deben no solo alcanzar una alta precisión en la identificación de objetivos, sino también garantizar la privacidad y seguridad de los datos, salvaguardando información sensible y manteniendo su operatividad en entornos dinámicos y adversarios. La resiliencia frente a ataques hostiles que buscan comprometer la integridad del modelo o interrumpir las operaciones de reconocimiento es esencial para el éxito de misiones en escenarios de alta amenaza.

El aprendizaje federado (FL), propuesto por Google en 2016, aborda algunos de estos desafíos mediante el entrenamiento de modelos de aprendizaje automático colaborativos y que preservan la privacidad. En este enfoque cada participante entrena un modelo localmente con sus propios datos y envía únicamente las actualizaciones del modelo (pesos o gradientes) a un servidor central, donde se agregan para construir un modelo global que posteriormente se redistribuye [1]. Este proceso iterativo permite la colaboración en el entrenamiento sin compartir directamente los datos, lo que protege la privacidad. En el campo del reconocimiento militar, FL se ha utilizado en vehículos aéreos no tripulados (UAVs) para aprender de forma colaborativa y mejorar los algoritmos de detección de enemigos, manteniendo la confidencialidad de los datos locales [6]. Sin embargo, las soluciones tradicionales basadas en FL dependen de un servidor central para la coordinación del entrenamiento, lo que introduce puntos únicos de fallo y cuellos de botella en la sincronización durante la agregación de modelos [7]. En misiones de reconocimiento aéreo, la dependencia de un servidor central es problemática, ya que las comunicaciones seguras y fiables pueden verse interrumpidas o incluso ser inexistentes en un entorno de combate. Además, la alta movilidad de los vehículos aéreos complica la conexión constante con una entidad centralizada.

Para abordar estas limitaciones, DFL [1] permite que cada participante entrene modelos locales y contribuya al modelo global mediante la agregación descentralizada. Los participantes comparten las actualizaciones directamente entre ellos, lo que garantiza una colaboración continua en ausencia de un servidor. La descentralización es particularmente ventajosa para las misiones militares de reconocimiento aéreo que operan en entornos altamente dinámicos caracterizados por maniobras complejas y el riesgo siempre presente de interferencias adversarias [8]. Para maximizar la eficacia de DFL en estas misio-

nes, los mecanismos de defensa robustos son esenciales para asegurar un intercambio de información fiable y seguro. Dado que DFL implica el intercambio directo de modelos entre los participantes, las técnicas de preservación de la privacidad son cruciales para proteger los datos sensibles y minimizar el riesgo de fuga de privacidad durante los intercambios. Estos mecanismos se implementan mediante la verificación de las actualizaciones de los modelos, el comportamiento de la comunicación y el cumplimiento de las rutas de vuelo designadas [9]. Al integrar la conciencia situacional en DFL, los mecanismos de defensa anteriores se refuerzan aún más, permitiendo la toma de decisiones informadas en tiempo real basadas en una comprensión global del campo de batalla [10]. Así, los vehículos militares pueden percibir, comprender y proyectar con precisión su entorno, lo que resulta crucial para la detección eficaz del enemigo y el éxito general de la misión. En las misiones aéreas militares, la conciencia situacional implica capacidades clave como la formación del vuelo, el geoposicionamiento, el uso de recursos y las interconexiones medias de los vehículos [5]. Cada capacidad garantiza que los vehículos aéreos mantengan las posiciones designadas, optimicen los recursos operativos y se comuniquen de forma fiable dentro del escuadrón, mejorando así la solidez y la seguridad de la federación.

A pesar de las ventajas de integrar mecanismos de defensa y la conciencia situacional con DFL, la literatura existente sólo ha abordado parcialmente estos retos. Algunos estudios han propuesto técnicas como blockchain para la descentralización y los registros seguros de transacciones [11], así como la implementación de algoritmos avanzados de agregación para mitigar las contribuciones maliciosas [9]. Sin embargo, los enfoques anteriores suelen tener limitaciones, como basarse en topologías de red estáticas o carecer de descentralización en funciones clave como la agregación de actualizaciones de modelos [6]. Además, no tienen plenamente en cuenta la naturaleza altamente dinámica de las misiones aéreas militares, lo que provoca interrupciones en las comunicaciones y actualizaciones incoherentes en DFL [10]. Muchas de las soluciones existentes también se centran en tipos de ataque específicos sin adoptar una estrategia de seguridad global y adaptable [8]. Aunque algunas investigaciones han explorado la aplicabilidad de la FL en contextos militares, siguen existiendo retos significativos en cuanto a la conciencia situacional, la convergencia de modelos y las amenazas adversarias. Para mejorar estos retos, este trabajo presenta:

- Un modelo de amenazas adaptado a los retos de DFL en escenarios dinámicos. Este modelo tiene en cuenta las misiones aéreas, aborda las interrupciones de las comunicaciones y los ciberataques adversarios que podrían afectar a las operaciones DFL y a la detección oportuna de vehículos enemigos.
- NEBULA-SA, una extensión de la plataforma NEBULA enfocada en misiones militares de reconocimiento aéreo [12]. Esta solución integra un módulo de defensa novedoso basado en la conciencia situacional para garantizar un entrenamiento seguro y descentralizado de los modelos federados, abordando eficazmente los desafíos únicos de los entornos militares dinámicos. Además, mejora la resiliencia de los modelos mitigando los ataques y manteniendo el rendimiento en condiciones adversas.
- Un escenario militar simulado con cuatro drones equipadas con SAR en diferentes formaciones de vuelo para detectar vehículos enemigos. Las extensas simulaciones, que incluían desviación del geoposicionamiento, manipulación del rumbo de colisión y ataques adversarios de envenenamiento de datos, mostraron i) una mitigación eficaz de los ataques a las comunicaciones y una mayor robustez del modelo, logrando una puntuación F1 de hasta el 95.8 %; y ii) un uso eficiente de los recursos y tiempos de respuesta, con una CPU en torno al 57 %, una RAM de 1.7 GB, y una reducción de la pérdida de paquetes manteniendo tiempos de ronda estables.

II. Trabajo Relacionado

FL ha sido explorado en diversos contextos militares, especialmente en operaciones de reconocimiento aéreo con vehículos autónomos. Sin embargo, la mayoría de las propuestas se han centrado en arquitecturas centralizadas, las cuales presentan vulnerabilidades como la dependencia de un servidor central. Por otro lado, enfoques descentralizados han comenzado a emerger para mejorar la resiliencia y autonomía de los UAVs en entornos hostiles. Tabla I resume los trabajos identificados desde distintas perspectivas: arquitectura de FL, métricas evaluadas, ataques abordados y soluciones propuestas para mejorar la seguridad y robustez del modelo.

Algunos trabajos recientes han demostrado cómo FL puede mejorar la autonomía de los UAVs en misiones de reconocimiento, aprovechando el aprendizaje distribuido para fortalecer la detección de objetivos sin comprometer la privacidad. Khalil & Rahman [6] presentan Fed-UP, una solución que combina FL con aprendizaje por refuerzo profundo para optimizar rutas de UAV en entornos hostiles. Su enfoque permite a los drones aprender colectivamente a evitar defensas antiaéreas, adaptándose dinámicamente a entornos en constante cambio. Este modelo es especialmente útil en situaciones donde la conectividad con un centro de comando es limitada y los UAVs deben tomar decisiones autónomas para mejorar la supervivencia en misiones de reconocimiento. En cuanto a la seguridad de los datos en FL, Lin & Zhang [7] proponen PPVerifier, un método verificable y preservador de privacidad basado en técnicas criptográficas. Su enfoque garantiza la integridad del modelo al evitar manipulaciones en las actualizaciones y ataques de privacidad durante la federación usando el conocimiento global que tiene el servidor central.

Otros estudios han explorado la descentralización como alternativa para mejorar la resiliencia. Qu et al. [8] introducen una solución de FL descentralizado para UAVs, centrándose en la optimización de comunicación y coordinación en enjambres aéreos. El objetivo es la optimización de rutas, adaptación del movimiento y resiliencia a los factores del entorno. Mientras que Cheriguene et al. [9] abordan el problema de envenenamiento de modelos mediante un mecanismo de selección de clientes confiables, que evalúa la reputación de cada UAV y excluye aquellos que presentan comportamientos maliciosos. Los autores únicamente emplean la similitud entre modelos para estimar la reputación asignada. Este método emplea agregaciones seguras orquestadas por un participante central que tiene visión del resto de participantes, permitiendo detectar patrones extraños en las interacciones durante la

Tabla I
COMPARACIÓN DE TRABAJOS PARA LA DETECCIÓN E IDENTIFICACIÓN DE VEHÍCULOS EN ESCENARIOS MILITARES USANDO FL

Referencia	Arquitectura FL	Métricas	Ataques (si aplica)	Objetivo / Solución Propuesta
[6]	Centralizado	Precisión	No	• FL + RL para planificación de ruta UAVs • Evita defensas antiaéreas
[7]	Centralizado	Verificabilidad, Privacidad	Sí (ataques de privacidad)	• FL verificable con técnicas criptográficas • Protección contra manipulación de modelos
[8]	Descentralizado	Precisión, Comunicación	No	• Arquitectura descentralizada para UAVs • Optimización de comunicación en enjambres
[9]	Centralizado	Precisión, Convergencia	Sí (envenenamiento de modelo)	• Selección de clientes confiables • Filtrado de UAVs maliciosos
[11]	Descentralizado	Seguridad, Integridad	Sí (seguridad, descentralización)	• Integración de blockchain en FL • Registro seguro de actualizaciones de modelo
NEBULA-SA (este trabajo)	**Descentralizado**	**Precisión, Robustez, Seguridad**	**Sí (ataques de datos y comunicación)**	• Federaciones totalmente descentralizadas y asíncronas • Conciencia situacional para detección de anomalías • Mayor resiliencia ante interrupciones de enlace • Protección contra envenenamiento de modelos

federación. Otra línea de trabajo se enfoca en la integración de tecnologías complementarias para reforzar la seguridad, Feng et al. [11] presentan un enfoque novedoso basado en blockchain para descentralizar FL en UAVs, asegurando que todas las actualizaciones del modelo sean verificables y resistentes a manipulaciones. La combinación de blockchain con FL descentralizado proporciona una mayor confianza en los modelos entrenados, reduciendo la posibilidad de ataques basados en modificaciones maliciosas de los parámetros del modelo.

En comparación con los enfoques anteriores, NEBULA-SA introduce una solución completamente descentralizada y asíncrona para FL en misiones de reconocimiento aéreo. A diferencia de las arquitecturas centralizadas, elimina la dependencia de un servidor único, evitando puntos de fallo y mejorando la resiliencia del sistema ante interrupciones de comunicación. Además, integra un módulo especializado en conciencia situacional, que monitoriza continuamente el entrenamiento y las interacciones físicas entre los participantes de la federación. De esta manera, permite detectar anomalías en las actualizaciones de los modelos y en los intercambios de información, mitigando intentos de envenenamiento de datos y garantizando la integridad del proceso federado.

III. MODELO DE AMENAZAS EN MISIONES MILITARES DE RECONOCIMIENTO BASADAS EN FL

Esta sección presenta las amenazas encontradas en misiones de reconocimiento asistidas por DFL, consideradas las más críticas debido a su impacto potencial en el éxito de la misión, como se ilustra en la Figura 1.

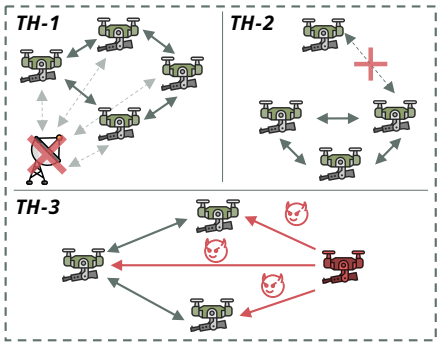


Figura 1. Amenazas en misiones de reconocimiento basadas en FL

TH-1: Debilidades de mando y control. Los vehículos aéreos militares dependen de la comunicación descentralizada para gestionar y sincronizar las actualizaciones del modelo en lugar de un servidor de mando central. Esta descentralización puede causar incoherencias en el rendimiento del modelo y problemas de sincronización durante las misiones [7]. La ausencia de un servidor central hace que sea difícil detectar y mitigar estos problemas con prontitud, lo que permite que un solo vehículo pueda poner en peligro toda la eficacia operativa. En un escenario militar de alto riesgo, podrían fracasar las misiones de reconocimiento y disminuir los servicios de inteligencia.

TH-2: Peligros para la movilidad operativa. Los vehículos aéreos en formación cambian constantemente de posición, lo que complica el mantenimiento de unas comunicaciones seguras y fiables [6]. Los adversarios podrían aprovechar los movimientos para interferir o secuestrar los canales de comunicación. Por ejemplo, un adversario que manipule la trayectoria de un vehículo aéreo mediante ciberataques a sus sistemas de navegación podría obligarlo a desviarse de su trayectoria prevista [10]. Estas desviaciones podrían provocar colisiones físicas, aislar el vehículo aéreo de la red o comprometer la federación de todo el escuadrón.

TH-3: Sabotaje de Vehículos Aéreos. Los vehículos aéreos dentro de la red pueden verse comprometidos o actuar maliciosamente, lanzando ataques adversarios como el envenenamiento de datos, en el que se inyectan datos falsos para degradar el rendimiento del modelo. Los vehículos aéreos maliciosos también pueden manipular las entradas para engañar al algoritmo de aprendizaje, lo que da lugar a predicciones incorrectas. Estas actividades merman considerablemente la eficacia del sistema, comprometiendo el éxito de la misión al hacer que el modelo pase por alto posiciones enemigas o identifique erróneamente unidades amigas como hostiles, lo que repercute directamente en las operaciones militares [9].

IV. DISEÑO E IMPLEMENTACIÓN DE NEBULA-SA

NEBULA-SA mejora las limitaciones de las soluciones FL existentes para misiones de reconocimiento aéreo militar incorporando DFL y un módulo avanzado de defensa basado en la conciencia situacional. Contrarresta eficazmente TH-1, TH-2 y TH-3 (véase la Sección III). La solución aprovecha la descentralización para mejorar la resistencia y la eficacia operativa, permitiendo que los vehículos aéreos operen de forma independiente a la vez que entrenan modelos de forma

colaborativa y segura. NEBULA-SA consta de los siguientes módulos (véase Figura 2): módulos de entrenamiento, comunicaciones y defensa, junto con dos módulos adicionales para simulación de escenarios y generación de amenazas.

El núcleo de NEBULA-SA está formado por los módulos de entrenamiento, comunicaciones y defensa, que gestionan la formación de modelos locales, el intercambio de datos y la protección del sistema, respectivamente. Inicialmente, cada vehículo aéreo recoge y preprocesa los datos de sus sensores a bordo, como los datos SAR para la detección de objetivos ①, y se dedica a la formación de modelos locales ②③. Una vez entrenados los modelos locales, los vehículos aéreos comparten periódicamente las actualizaciones de sus modelos con sus vecinos ④. Para proteger los datos sensibles, NEBULA-SA integra privacidad diferencial (DP) adaptativa, que aplica selectivamente restricciones de privacidad reteniendo los parámetros críticos y añadiendo ruido a los menos significativos [15]. Esto mejora la privacidad sin comprometer el rendimiento o la convergencia del modelo. DP combina estrategias de restricción diferencial y un *clip norm decay* para defenderse contra la fuga de privacidad, el envenenamiento y los ataques de puerta trasera. El *clip norm decay* reduce el ruido inyectado disminuyendo gradualmente el umbral de recorte de las actualizaciones del modelo a lo largo del entrenamiento, preservando la utilidad al tiempo que mejora la protección. Además, NEBULA-SA incorpora un sofisticado módulo de comunicaciones que utiliza un mecanismo de comunicación asíncrona tolerante a fallos, compatible con diversas topologías de red y participantes móviles. Aprovecha la transmisión de datos por trozos adaptada a las condiciones de la red con trazabilidad de la pérdida de paquetes. El módulo de comunicación utiliza compresión LZ4, lo que demuestra un equilibrio óptimo entre velocidad y eficiencia [14]. Los mensajes de descubrimiento restablecen dinámicamente las conexiones a medida que los vehículos aéreos se desplazan, lo que mejora la gestión de la movilidad. Si se producen desconexiones debido a tiempos de respuesta lentos o a un rendimiento degradado, NEBULA-SA inicia de forma proactiva las reconexiones para mantener una integración sin fisuras en la red. Además, un mecanismo de congestión redirige las comunicaciones por rutas alternativas para optimizar el rendimiento y la fiabilidad. NEBULA-SA modela el impacto en la latencia de las comunicaciones en lugar del movimiento físico para simular la movilidad. Los cambios en las posiciones relativas de los vehículos aéreos influyen en la estabilidad de las comunicaciones, con variaciones en la latencia correspondientes a las distancias entre nodos.

El módulo de defensa mitiga el riesgo que suponen los nodos maliciosos que intentan introducir modelos envenenados o adoptar comportamientos adversos dentro de la federación. Esto se consigue evaluando criterios clave, entre los que se incluyen:

- *Semejanza de modelos*. Cada vehículo envía periódicamente actualizaciones del modelo $\mathbf{w}_i^t$ en la ronda t de la federación. El módulo de defensa realiza una comprobación de similitud del coseno para comparar estas actualizaciones con el modelo agregado de la ronda anterior $\mathbf{w}^t$ y las actualizaciones de otros participantes $\frac{\mathbf{w}_i^t \cdot \mathbf{w}^{t-1}}{\|\mathbf{w}_i^t\| \|\mathbf{w}^{t-1}\|}$, donde $\mathbf{w}_i^t \cdot \mathbf{w}^{t-1}$ es el producto escalar entre los vectores de pesos. Adicionalmente, el módulo analiza la desviación acumulada de parámetros, definida como: $D_i^t = \|\mathbf{w}_i^t - \mathbf{w}^{t-1}\|$, siendo D_i^t la distancia euclidiana entre la actualización enviada y el modelo de referencia. Se establece un umbral adaptativo Θ^t tal que si $D_i^t > \Theta^t$, la actualización se considera sospechosa.
- *Conciencia de la situación*. El módulo evalúa la integridad operativa de cada vehículo utilizando diversas métricas de la conciencia situacional, incluyendo:
  - *Formación de vuelo*. El módulo supervisa la adherencia de cada vehículo aéreo a la formación designada comparando su posición actual $\mathbf{p}_i(t)$ con la posición esperada $\mathbf{p}_i^{\text{ref}}(t)$. La desviación de formación se cuantifica como: $D_f^t = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i(t) - \mathbf{p}_i^{\text{ref}}(t)\|$ donde N es el número de vehículos en la formación. Un valor alto de D_f^t puede indicar una pérdida de cohesión en la formación, afectando la operatividad del escuadrón y elevando el riesgo de detección por fuerzas adversarias.
  - *Geoposicionamiento*. Para evaluar posibles manipulaciones en la navegación, se calcula la distancia euclidiana entre la posición actual $\mathbf{p}_i(t)$ y la trayectoria planificada $\mathbf{p}_i^{\text{traj}}(t)$: $D_g^t = \|\mathbf{p}_i(t) - \mathbf{p}_i^{\text{traj}}(t)\|$.
  - *Interconexión*. La fiabilidad de las comunicaciones se mide a través de la latencia media de los enlaces de un nodo i respecto a sus vecinos: $L_i^t = \frac{1}{|N_i|} \sum_{j \in N_i} L_{ij}^t$ donde L_{ij}^t es la latencia del enlace entre los vehículos aéreos i y j. Un aumento en la variabilidad $\Delta L_i^t = |L_i^t - L_i^{t-1}|$ puede indicar intentos de interferencia en la red o fallos en la conectividad.
  - *Uso de Recursos*. El módulo realiza un seguimiento de los recursos consumidos por cada vehículo aéreo, por ejemplo, $U_{\text{CPU}}^t = \frac{C_{\text{uso}}^t}{C_{\text{total}}}$, $U_{\text{RAM}}^t = \frac{M_{\text{uso}}^t}{M_{\text{total}}}$. Un uso de recursos inusualmente alto o bajo puede indicar actividad maliciosa o intentos de interferir en el proceso de entrenamiento.

Tras recopilar las métricas, se calcula una puntuación de defensa ⑤ para cada vehículo aéreo mediante un proceso de tres pasos (véase Figura 3): Ⓐ las métricas se estandarizan en una escala común $[0-1]$ para mantener la coherencia; Ⓑ se aplican ponderaciones a cada métrica, siendo la similitud del modelo la que recibe la mayor prioridad (30 % de la puntuación total) debido a su papel fundamental en la detección de posibles ataques de envenenamiento, sin comprometer el rendimiento en federaciones con datos heterogéneos. El 70 % restante se distribuye entre las métricas de la conciencia situacional en función de su importancia para el contexto operativo: formación de vuelo (30 %), geoposicionamiento (15 %), interconexión (15 %) y uso de recursos (10 %); iii) se calcula una suma ponderada de estas métricas, lo que da como resultado la puntuación de defensa Ⓒ, que refleja la fiabilidad general del vehículo dentro de la federación. Estas ponderaciones pueden ajustarse al caso de uso específico y a los requisitos operativos.

A continuación, esta puntuación de defensa se compara con un umbral dinámico que se adapta en función de las condiciones de entrenamiento

$$\Theta(t) = \frac{0.9}{1 + e^{-k(t-0.2T)}}, \tag{1}$$

donde T es el número total de rondas federadas y k controla

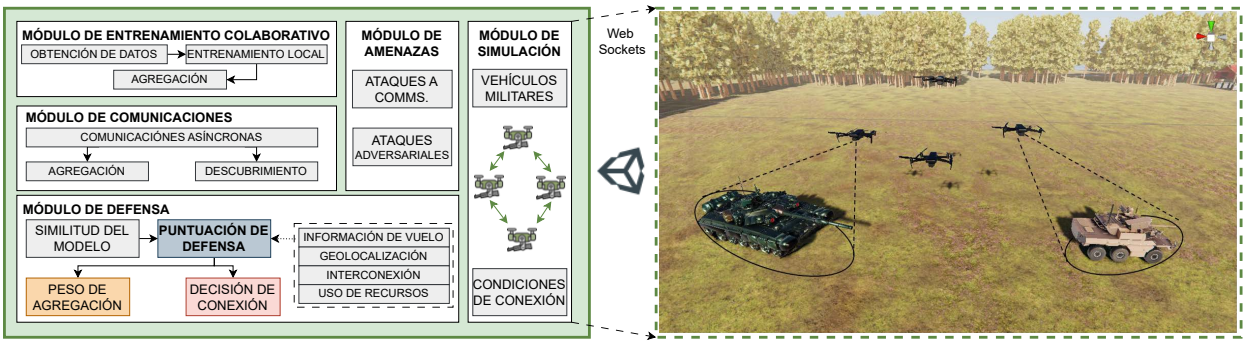


Figura 2. Visión general de NEBULA-SA

la transición del umbral. . En las primeras rondas (el primer 20 % de las rondas federadas), las actualizaciones del modelo son más variables debido al ruido y la incertidumbre. Para evitar la detección errónea de vehículos benignos, el término de regularización $\Omega_i(t)$ atenúa estas fluctuaciones

$$\Omega_i(t) = \sum_{k=1}^{6} \sigma_k^2 \cdot (\mathcal{M}_{i,k}(t) - \mathbb{E}[\mathcal{M}_k(t)])^2, \qquad (2)$$

donde: $\mathcal{M}_{i,k}(t)$ es el valor de la métrica k para el nodo i en la ronda t, $\mathbb{E}[\mathcal{M}_k(t)]$ representa la media esperada de la métrica k en la federación, y σ_k^2 es la varianza de la métrica k. A medida que el modelo converge, el efecto de la regularización se intensifica, penalizando más las desviaciones inesperadas. Como resultado, el umbral $\Theta(t)$ aumenta gradualmente hasta 0.9, asegurando que las desviaciones en las etapas finales del entrenamiento tengan una mayor probabilidad de indicar comportamientos maliciosos. Los vehículos con puntuaciones de defensa por encima del umbral dinámico se clasifican como benignos, recibiendo un mayor peso en el proceso de agregación para priorizar las entradas de confianza ⑥. Los que están por debajo del umbral se marcan como potencialmente maliciosos, y su peso de contribución en la agregación se reduce. Si este comportamiento continúa durante tres o más rondas, sus conexiones se cortan para proteger la integridad del proceso de aprendizaje ⑦. Figura 3 muestra el flujo operacional completo y la interacción entre los módulos de entrenamiento, comunicación y defensa.

Finalmente, los módulos de simulación y de amenazas son esenciales para evaluar NEBULA-SA, estableciendo el escenario del campo de batalla. El módulo de simulación permite crear escenarios militares aéreos realistas y dinámicos. Despliega vehículos aéreos militares simulados utilizando contenedores Docker, emulando el comportamiento de drones individuales. La simulación también abarca las condiciones de la red y la ruta, proporcionando un entorno de red completo que modela los parámetros de ancho de banda, retardo y pérdida. Además, gestiona patrones de formación militar estándar como Diamond y Right Hand Finger Four [16], con distancias estandarizadas entre vehículos aéreos para garantizar una comunicación y coordinación óptimas. Al mismo tiempo, la simulación ofrece una representación en tiempo real mediante el motor gráfico Unity, que visualiza las trayectorias de vuelo y provee características realistas sobre geoposicionamiento. El módulo de amenazas puede

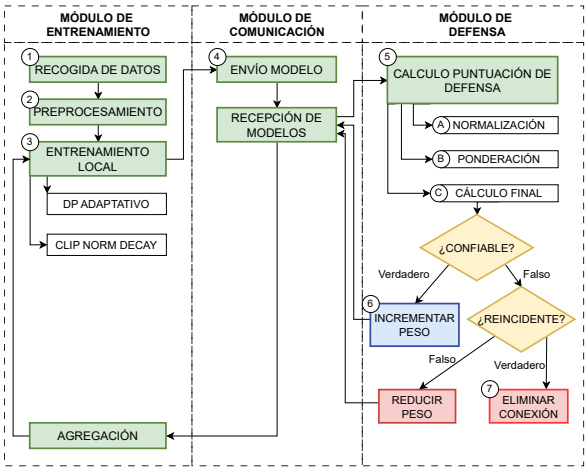


Figura 3. Flujo operacional de NEBULA-SA

introducir perturbaciones en la longitud, latitud y altitud de un vehículo aéreo, haciendo que se desvíe de la formación o se acerque a otros vehículos. El módulo también simula ataques adversarios al modelo federado inyectando información falsa en los modelos locales, imitando el envenenamiento de datos y los ataques a la integridad. Estos ataques se basan en el modelo de amenaza de la Sección III.

V. Casos de Uso

Se han definido cuatro misiones para reproducir las amenazas descritas en la Sección III: una misión de reconocimiento sin ataques y tres con diferentes ataques (desviación del geoposicionamiento, manipulación del rumbo de colisión y envenenamiento de datos por el adversario). Los ataques fueron generados por un adversario cuyo objetivo era un dron en la ronda 20, desplegada a través del módulo de amenazas de NEBULA-SA.

V-A. Misión de Reconocimiento

Se desplegó una simulación de cuatro drones utilizando contenedores Docker, que ofrecían una solución ideal para replicar la interacción entre drones gracias a su capacidad para garantizar el aislamiento de recursos, la comunicación entre nodos y la gestión eficiente de los recursos del sistema. La plataforma utilizada para la orquestación de las entidades

en la federación fue NEBULA, la cual coordina el ciclo federado mediante una arquitectura descentralizada basada en mensajes punto a punto, sin depender de un servidor central [12]. El sistema host tenía una CPU Intel i7-10700F y 100 GB de RAM, y a cada dron se le asignó un núcleo diferente para imitar las capacidades de procesamiento realistas. La flexibilidad y escalabilidad de Docker permitieron una integración perfecta de los componentes clave de la simulación, como el seguimiento de la geolocalización, las rutas de movimiento y los escenarios de ataque, garantizando un modelado preciso de las complejas operaciones aéreas y las interacciones adversarias. Para mejorar la representación del entorno operativo, se integró Unity en paralelo con el despliegue en Docker, estableciendo una interfaz abierta y bidireccional entre NEBULA-SA y Unity mediante el uso de WebSockets. A través de esta interfaz, se recopilan métricas clave de conciencia situacional, tales como condiciones de distancia entre drones, posicionamiento relativo, formaciones de vuelo y la interconexión entre los vehículos aéreos. Las condiciones de la red se simularon para reflejar los retos de la comunicación aérea. A cada dron se le asignó un ancho de banda de 3 MB/s utilizando la envoltura *tc* en la interfaz de comunicación. Además, se introdujeron retardos de red de 5-10 ms para reproducir el intercambio real de mensajes en el aire.

Se equipó a aviones militares con un sistema SAR simulado para detectar vehículos militares enemigos. La tecnología SAR proporciona imágenes de teledetección de alta resolución independientemente del tiempo, la hora o las condiciones de luz. La misión utilizó tres conjuntos de datos para evaluar la eficacia de NEBULA-SA. En primer lugar, el conjunto de datos MSTAR, desarrollado por el Laboratorio de Investigación de las Fuerzas Aéreas y DARPA, incluye imágenes SAR de diversos vehículos militares, como tanques y vehículos de transporte de tropas. En segundo lugar, el conjunto de datos SAMPLE proporciona pares de imágenes medidas y sintéticas para diez clases de objetivos MSTAR públicos, ampliando la diversidad de entrenamiento y evaluación. Por último, el conjunto de datos OpenSARShip, compuesto por chips de buques SAR recogidos a partir de 41 imágenes Sentinel-1 de banda C, contiene tipos de buques como petroleros, portacontenedores y graneleros para evaluar el rendimiento en escenarios marítimos.

Los datos se distribuyeron como No Independientes e Idénticamente Distribuidos (Non-IID) utilizando una distribución Dirichlet con $\alpha = 0.3$ y $\alpha = 0.9$, simulando distintos niveles de heterogeneidad de datos entre drones. Con $\alpha = 0.3$, los datos imitan las condiciones del mundo real, en el que algunos drones detectan principalmente clases específicas, lo que provoca que alrededor del 20 % de las clases estén poco representadas. En cambio, $\alpha = 0.9$ ofrece una distribución de clases más equilibrada y más adecuada para contextos operativos uniformes. Para el resto de los experimentos, se eligió $\alpha = 0.3$ por su mayor adecuación a escenarios de datos heterogéneos. Cada avión simulado desplegó una arquitectura VGG16, que presenta una red neuronal profunda con 16 capas, compuestas principalmente de capas convolucionales, seguidas de capas totalmente conectadas para la clasificación. Los modelos se entrenaron a lo largo de 20 épocas utilizando el método de Descenso Gradiente Estocástico (SGD) para ajustar los hiperparámetros.

Durante la misión de reconocimiento, las formaciones de aviones se alternaban cada 5 minutos entre Diamond and Right Hand Finger Four, manteniendo separaciones laterales de 1 milla y verticales de 100 pies. Cada dron comenzó con una posición estandarizada relativa al esquema de la formación. El módulo de simulación controlaba continuamente estas distancias para garantizar un comportamiento correcto de la formación, que era detectado como normal por las métricas de la conciencia situacional. El proceso de federación implicó 50 rondas. Este número de rondas se eligió en función de nuestros experimentos, que demostraron una convergencia óptima del modelo en ese punto. Tabla II detalla la configuración y las especificaciones de los drones militares desplegadas durante la misión.

Tabla II
CONFIGURACIÓN DE LOS AVIONES MILITARES DURANTE LAS MISIONES

Característica	**Descripción**
Nodos de la Federación	• Cuatro aviones militares simulados (contenedores Docker) • Equipados con NEBULA-SA
Simulación	• Uso de UNITY 2020.3.30f1 (campo de batalla realista) • Obtención de métricas de geolocalización e interconexión
Formación aérea	• Diamond and Right Hand Finger Four • 1 milla lateralmente, 100 pies verticalmente • Movimiento dinámico cada 5 min
Dataset	• Datos non-IID (Dirichlet $\alpha = 0.3$) • MSTAR y SAMPLE con diez vehículos militares terrestres • OpenSARShip con seis vehículos marítimos
Modelo federado	• VGG16 • 50 rondas federadas
Comunicaciones	• 2-3 MB/s de ancho de banda con una pérdida de mensajes del 10 % • Retraso de 5-10 ms • Transmite datos de geoposicionamiento diez veces por segundo
Ataques	① 10-15° de latitud/longitud, 50-200 metros de altitud ② 200-250 metros hacia sus vecinos (gradual) ③ - Carlini & Wagner (0.9 C, 0.01 α, 1000 iteraciones) - Gradient deviation (basado en la media) [17] - Backdoor pixel trigger (arriba-izquierda "X", objetivo aleatorio)

V-B. Ataque I: Desviación del geoposicionamiento

En este escenario, un atacante externo afecta al módulo de geoposicionamiento de un dron introduciendo perturbaciones en sus datos de latitud, longitud y altitud, alineándose con TH-1. Esto provocó que el dron se desviara de su trayectoria de vuelo prevista, provocando su aislamiento de la formación y comprometiendo la coordinación y comunicación con otros drones. Desde la perspectiva DFL, este aislamiento significa que el conocimiento y los datos de los drones afectados, cada una de las cuales posee subconjuntos diferentes del conjunto de datos, se excluyeron de las actualizaciones del modelo. Esta exclusión da lugar a tiempos de ronda más largos debido a las prolongadas esperas en la recepción de datos y reduce la precisión global del proceso de aprendizaje. En consecuencia, el ataque perturbó la formación física y degradó la inteligencia colectiva y la eficacia operativa de todo el escuadrón, comprometiendo el éxito y la resistencia de la misión.

V-C. Ataque II: Manipulación del rumbo de colisión

En el segundo escenario, el ataque manipuló los datos de la trayectoria de un dron para acercarla peligrosamente a otros drones, alineándose con TH-2. Al introducir desviaciones que alteraban el rumbo del dron entre 200 y 250 metros hacia sus vecinos, el ataque creaba un rumbo de colisión. Esta manipulación aumentaba el riesgo de colisiones en el aire

y perturbaba la cohesión de la formación, amenazando la seguridad de los drones y con el objetivo de crear confusión y obstaculizar la eficacia operativa. Desde la perspectiva DFL, las maniobras erráticas o las acciones evasivas repentinas reducirían la fiabilidad de las actualizaciones del modelo proporcionadas por los drones afectados. Además, ajustar las formaciones o aumentar las distancias entre drones para acomodar estos movimientos alargaría el tiempo de convergencia debido al aumento de la latencia de las comunicaciones.

V-D. Ataque III: Envenenamiento adversarial de datos

El tercer escenario incluía tres escenarios de ataque adversario distintos para evaluar la robustez del módulo de defensa, alineado con TH-3. *ADP-1* empleó un ataque de evasión inspirado en el método Carlini & Wagner. Este ataque inyectaba sutiles perturbaciones en los datos de las imágenes SAR utilizando una puntuación de confianza de 0.9, una tasa de aprendizaje de 0.01 y hasta 1000 iteraciones, lo que daba lugar a clasificaciones erróneas, como la identificación de un tanque enemigo como un camión de evacuación médica, degradando así el rendimiento del modelo federado. Los parámetros se seleccionaron empíricamente para lograr una reducción no selectiva de más del 50 % en el rendimiento del modelo con una perturbación mínima. *ADP-2* se basó en [17], que manipula gradientes introduciendo desviaciones controladas derivadas de la media y la desviación estándar de actualizaciones benignas, permitiendo al adversario influir en el modelo dentro de límites de parámetros plausibles y sin acceso completo a los gradientes. *ADP-3* implicó un ataque de puerta trasera, en el que se introducía un patrón de píxeles `X` en la esquina superior izquierda de imágenes objetivo aleatorias. Este disparador incrustado permite a los modelos maliciosos clasificar erróneamente las imágenes que contienen el disparador (por ejemplo, vehículos enemigos) mientras que por lo demás se comportan normalmente, lo que dificulta su detección mediante métodos de seguimiento convencionales.

VI. Resultados

En la Tabla III se compara el rendimiento de NEBULA-SA en cuatro escenarios anteriores (normal y bajo ataque) utilizando el modelo VGG16 y los conjuntos de datos de la Sección V-A. Además, se compara con una solución de referencia en DFL, como Fedstellar [13], que refleja el proceso de formación, la arquitectura del modelo, el procesamiento de datos y las rondas de federación de NEBULA-SA. Sin embargo, la solución de referencia en DFL carece del módulo de defensa basado en la conciencia situacional que permite a NEBULA-SA realizar ajustes de agregación adaptativos y responder en tiempo real a los ciberataques.

En la misión básica de reconocimiento sin ataques, NEBULA-SA logró un alto rendimiento del modelo, con F1-score del 97.5 % en los conjuntos de datos MSTAR y SAMPLE y del 79.1 % en OpenSARShip. El módulo de defensa introdujo una sobrecarga mínima, con un uso de la CPU que pasó del 45 % en el framework de referencia al 57 % en NEBULA-SA. El impacto sobre el ancho de banda de la red y los tiempos de ronda fue insignificante, lo que indica que el módulo de defensa proporcionado por NEBULA-SA no compromete el rendimiento en condiciones normales.

Durante las misiones bajo ataques, NEBULA-SA contrarrestó con éxito los tres ataques definidos. En el escenario de desviación del geoposicionamiento, mantuvo puntuaciones F1 en torno al 91 % en MSTAR y SAMPLE y del 72.0 % en OpenSARShip, con un uso de CPU que aumentó moderadamente hasta el 67 %. La eficiencia de la red se mantuvo estable, con una pérdida de paquetes inferior al 1 % y un ligero aumento de los tiempos de ronda. Por el contrario, el rendimiento de la solución de referencia disminuyó significativamente, con puntuaciones F1 que cayeron al 70.2 % en MSTAR y SAMPLE y al 55.1 % en OpenSARShip, lo que pone de manifiesto su incapacidad para mitigar el impacto del dron comprometido.

Bajo el ataque de manipulación del rumbo de colisión, NEBULA-SA mantuvo puntuaciones F1-score por encima del 89 % en MSTAR y SAMPLE y del 70.1 % en OpenSARShip, identificando y aislando eficientemente el dron comprometido sin sobrecargar los recursos del sistema (uso de CPU del 65 %). Por el contrario, la solución de referencia sufrió una grave degradación, con puntuaciones F1 por debajo del 43.3 % en MSTAR y SAMPLE y del 30.1 % en OpenSARShip, incapaz de evitar colisiones o gestionar interrupciones de la comunicación. La pérdida de paquetes aumentó hasta casi el 5 %, y los tiempos de ronda se alargaron considerablemente.

Por último, en el ataque adversario de envenenamiento de datos, los mecanismos de defensa proporcionados por NEBULA-SA fueron especialmente eficaces. A pesar de los ataques, NEBULA-SA mantuvo las puntuaciones F1 entre el 52-90 %, dependiendo del tipo de ataque y del conjunto de datos, detectando y filtrando los datos maliciosos. En particular, en la desviación de gradiente y el ataque de puerta trasera en el conjunto de datos OpenSARShip, la puntuación F1 disminuyó a 52.1 % y 58.8 %, lo que indica un impacto significativo, pero sigue superando a la solución de referencia. El uso de recursos aumentó debido a las demandas computacionales del módulo de defensa, con un uso de la CPU que alcanzó el 75 %, aunque permaneció dentro de límites aceptables. La solución de referencia experimentó un drástico descenso en el rendimiento del modelo, con puntuaciones F1 que cayeron en picado hasta un 15 % bajo ciertos ataques como la desviación del gradiente. Esto pone de manifiesto su vulnerabilidad y el grave impacto de los datos maliciosos en el rendimiento del modelo sin una protección adecuada.

VII. Conclusiones

Este trabajo presenta NEBULA-SA, una extensión de la plataforma NEBULA que implementa un módulo de defensa basado en la conciencia situacional para mejorar el rendimiento de DFL en misiones militares bajo ataque. El módulo desarrollado asigna una puntuación de defensa a cada vehículo en función de la similitud del modelo y las métricas de la conciencia situacional, priorizando las entradas de fuentes fiables y mitigando las amenazas potenciales. NEBULA-SA se validó en una misión de reconocimiento simulada con cuatro aviones militares con sistemas SAR para identificar vehículos enemigos utilizando conjuntos de datos MSTAR, SAMPLE y OpenSarShip. La misión se enfrentó a desviaciones del geoposicionamiento que afectaban a la precisión de la localización, a la manipulación del rumbo de colisión que comprometía las trayectorias de vuelo y a ataques

Tabla III
MÉTRICAS DE RENDIMIENTO MEDIO DE NEBULA-SA PARA LA MISIÓN DE RECONOCIMIENTO BAJO DIFERENTES ATAQUES UTILIZANDO VGG16 A TRAVÉS DE TRES CONJUNTOS DE DATOS (MSTAR, SAMPLE, OPENSARSHIP)

Misión aérea militar	**Rendimiento del modelo (F1 score)**			**Uso de recursos**				**Tiempo (s)**
	MSTAR	**SAMPLE**	**OpenSARShip**	**CPU (%)**	**RAM (GB)**	**Ancho de banda (MB/s)**	**Pérdida de paquetes (%)**	
NEBULA-SA ***(CON** módulo de defensa)*								
Reconocimiento *(baseline)*	95.8±0.2	97.5±0.3	79.1±0.5	57±4	1.7±0.2	2.6±0.1	0.7±0.1	5.3±0.3
Desviación del geoposicionamiento	91.8±2.0	90.3±1.8	72.0±1.5	67±3	1.8±0.2	2.7±0.1	0.7±0.2	5.4±0.2
Manipulación del rumbo de colisión	92.5±1.9	89.8±2.1	70.1±2.2	65±2	1.7±0.2	2.5±0.2	0.8±0.3	5.4±0.1
Envenenamiento adversarial de datos								
ADP-1 (Carlini & Wagner)	90.1±2.5	88.7±2.3	65.0±1.9	67±4	2.2±0.2	2.8±0.2	0.7±0.1	5.6±0.3
ADP-2 (Gradient deviation [17])	71.3±2.9	72.9±2.0	52.1±1.4	75±3	2.2±0.3	2.9±0.2	0.8±0.1	5.6±0.5
ADP-3 (Backdoor pixel trigger)	85.0±2.0	83.5±1.7	58.8±1.9	73±8	2.2±0.2	2.9±0.2	0.9±0.1	5.6±0.3
Solución de referencia en DFL ***(SIN** módulo de defensa)*								
Reconocimiento *(baseline)*	98.1±0.2	95.5±0.3	78.1±0.5	45±3	1.5±0.2	2.5±0.1	0.7±0.1	5.2±0.3
Desviación del geoposicionamiento	72.3±2.5	70.2±2.3	55.1±1.5	61±3	1.8±0.2	2.7±0.1	0.9±0.2	5.4±0.3
Manipulación del rumbo de colisión†	43.3±1.6	41.0±1.7	30.1±1.2	65±2	1.7±0.2	2.5±0.2	0.9±0.6	5.4±0.1
Envenenamiento adversarial de datos								
ADP-1 (Carlini & Wagner)	18.5±2.3	17.5±2.0	16.3±1.9	70±8	2.2±0.2	2.8±0.2	1.1±0.6	5.6±0.3
ADP-2 (Gradient deviation [17])	15.0±2.5	18.9±4.7	12.0±2.4	75±4	2.8±0.2	2.4±1.2	0.8±0.1	6.1±0.7
ADP-3 (Backdoor pixel trigger)	25.1±3.4	22.3±2.8	19.8±2.5	86±7	3.4±0.6	2.9±1.2	0.9±0.1	5.9±0.6

adversarios de envenenamiento de datos dirigidos contra la integridad de los datos del modelo. Los resultados mostraron mejoras significativas con respecto a la línea de base, con puntuaciones F1 en MSTAR del 91.8 % durante la desviación del geoposicionamiento, del 92.5 % bajo la manipulación del rumbo de colisión y de hasta el 90.1 % al contrarrestar el envenenamiento de datos por parte de adversarios, mientras que el uso de recursos se mantuvo similar al de la misión sin ataques.

En futuros trabajos, se explorarán algoritmos de agregación avanzados para fortalecer la seguridad y eficiencia del entrenamiento descentralizado. Además, se evaluará NEBULA-SA en entornos físicos, desplegando el sistema en dispositivos embebidos y con restricciones de hardware, como Raspberry Pi, Jetson Nano y UAVs comerciales, validando su rendimiento en plataformas con recursos computacionales limitados. Este despliegue permitirá analizar la viabilidad del sistema en misiones reales, asegurando su adaptabilidad a escenarios operativos dinámicos. Asimismo, se investigará el impacto de condiciones de red reales, como enlaces de comunicación intermitentes, variabilidad del ancho de banda y latencias prolongadas en entornos de combate.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por *(a)* 21629/FPI/21, Fundación Séneca, Región de Murcia (Spain), *(b)* the European Commission through the Horizon Europe/JU SNS project ROBUST-6G (Grant Agreement no. 101139068), *(c)* DEFENDER from the Spanish National Institute of Cybersecurity (INCIBE) and by the Recovery, Transformation and Resilience Plan, Next Generation EU, y *(d)* the Swiss Federal Office for Defense Procurement (armasuisse) with DECIMAL project.

REFERENCIAS

[1] E. T. Martínez Beltrán, M. Quiles Pérez, P. M. Sánchez Sánchez, S. López Bernal, G. Bovet, M. Gil Pérez, G. Martínez Pérez, and A. Huertas Celdrán, "Decentralized Federated Learning: Fundamentals, State of the Art, Frameworks, Trends, and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 41, pp. 2983-3013, 2023.

[2] E. T. Martínez Beltrán, P. M. Sánchez Sánchez, S. López Bernal, G. Bovet, M. Gil Pérez, G. Martínez Pérez, and A. Huertas Celdrán, "Mitigating communications threats in decentralized federated learning through moving target defense," *Wireless Network*, 2024.

[3] C. Feng, A. Huertas Celdrán, J. Baltensperger, E. T. Martínez Beltrán, G. Bovet, and B. Stiller, "Sentinel: An Aggregation Function to Secure Decentralized Federated Learning," *arXiv preprint arXiv:2310.08097*, 2023.

[4] D. Qi, J. Zhang, X. Liang, Z. Li, J. Zuo, and P. Lei, "Autonomous Reconnaissance and Attack Test of UAV Swarm Based on Mosaic Warfare Thought," *International Conference on Robotics and Automation Engineering*, pp. 79-83, 2021.

[5] A. Sumari, K. Widyasari, and V. Lestari, "A Knowledge Growing System-based Decision Making-Support System Application for Forces Command and Control in Military Operations," *International Conference on ICT for Smart Society*, pp. 1-6, 2021.

[6] A. A. Khalil and M. A. Rahman, "FED-UP: Federated Deep Reinforcement Learning-based UAV Path Planning against Hostile Defense System," *International Conference on Network and Service Management*, pp. 268-274, 2022.

[7] L. Lin and X. Zhang, "PPVerifier: A Privacy-Preserving and Verifiable Federated Learning Method in Cloud-Edge Collaborative Computing Environment," *IEEE Internet of Things Journal*, pp. 8878-8892, 2023.

[8] Y. Qu, H. Dai, Y. Zhuang, J. Chen, C. Dong, F. Wu, and S. Guo, "Decentralized Federated Learning for UAV Networks: Architecture, Challenges, and Opportunities," *IEEE Network*, vol. 35, no. 6, pp. 156-162, 2021.

[9] Y. Cheriguene, W. Jaafar, H. Yanikomeroglu, and C. A. Kerrache, "Towards Reliable Participation in UAV-Enabled Federated Edge Learning on Non-IID Data," *IEEE Open Journal of Vehicular Technology*, vol. 5, pp. 125-141, 2023.

[10] B. Brik, A. Ksentini, and M. Bouaziz, "Federated Learning for UAVs-Enabled Wireless Networks: Use Cases, Challenges, and Open Problems," *IEEE Access*, vol. 8, pp. 53841-53849, 2020.

[11] C. Feng, B. Liu, K. Yu, S. K. Goudos, and S. Wan, "Blockchain-Empowered Decentralized Horizontal Federated Learning for 5G-Enabled UAVs," *IEEE Transactions on Industrial Informatics*, pp. 3582, 2022.

[12] E. T. Martínez Beltrán, "CyberDataLab/nebula," *GitHub Repository*, 2024. Available: https://github.com/CyberDataLab/nebula.

[13] E. T. Martínez Beltrán, Á. L. Perales Gómez, C. Feng, P. M. Sánchez Sánchez, S. López Bernal, G. Bovet, M. Gil Pérez, G. Martínez Pérez, and A. Huertas Celdrán, "Fedstellar: A Platform for Decentralized Federated Learning," *Expert Systems with Applications*, vol. 242, pp. 122861, 2024.

[14] A. Parthasarathy and B. Krishnamachari, "DEFER: Distributed Edge Inference for Deep Neural Networks," *International Conference on Communication Systems & Networks*, pp. 749, 2022.

[15] X. Yang, W. Huang, and M. Ye, "Dynamic Personalized Federated Learning with Adaptive Differential Privacy," *International Conference on Neural Information Processing Systems*, 2024.

[16] Formation and Safety Team (FAST), "Fundamentals of Formation for All Aircraft Based on FAST Formation Protocols," *FAST Reference Material*, 2016. Available: http://flyfast.org/content/reference-material.

[17] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning," *USENIX Conference on Security Symposium*, Article 92, 2020.

Toward a Military Smart Cyber Situational Awareness (CSA)

Anthony Feijóo-Añazco[1], Antonio López Martínez[2], Daniel Díaz-López[1,2]
Angel Luis Perales Gómez[3], Pantaleone Nespoli[1], Gregorio Martínez Pérez[1]
[1]Department of Information and Communications Engineering, University of Murcia, 30100, Murcia, Spain
{anthonyjosep.feijooa, antonio.lopezm2, danielorlando.diaz, pantaleone.nespoli, gregorio}@um.es
[2]School of Engineering, Science and Technology, Universidad del Rosario, Bogotá, Colombia
danielo.diaz@urosario.edu.co
[3]Department of Computers Engineering and Technology, University of Murcia, 30100, Murcia, Spain
angelluis.perales@um.es

***Abstract*—The development of technology across multiple sectors and the growing importance of cyber warfare make the development of Cyber Situational Awareness (CSA) a fundamental component of any cyber defense strategy. CSA, as a practice, enables understanding of the current landscape within an organization or critical infrastructure, anticipating potential threats, and responding appropriately to cyber risks. With CSA, we are not simply seeking a passive point of view, but rather informed decision-making that allows us to improve response times and monitor the consequences and effects an attack has on one of our elements and how it will affect other elements it interacts with. In this paper, we review 5 CSA platforms, seeking differentiating characteristics between each proposal and outlining 6 proposed criteria that can be applied when creating a military smart CSA platform. To this end, we have validated the proposed criteria in CRUSOE, an open-source CSA platform developed by CSIRT-MU. After applying some modifications and experiments, it turned out to be applicable to this field.**

***Index Terms*—Cyber Situational Awareness (CSA), Cyber Operational Picture (COP), Threat Intelligence, Cyber Defense**

Contribution type: *Original research*

I. Introduction

Situational Awareness (SA) arises from the military environment where the aim is to know the situation of different elements and estimate how a decision may affect them [1], [2]. In the cybersecurity domain, Cyber Situational Awareness (CSA) applies SA principles to cyberspace, focusing on real-time monitoring, analysis, and response to threats [3]. Specifically, CSA involves identifying the behavior of adversaries and how their attack may affect assets to prioritize strategies and ensure the successful course of operations [4].

CSA plays an important role in cyber defense due to its dynamic and predictive functionality, as well as real-time monitoring, allowing commanders to take proactive measures to allocate resources and validate system designs against threats. This capability uses Mission Engineering (ME), a methodology that aligns technical capabilities with operational objectives through threat impact modeling [5]. ME relies on CSA data to quantify how attacks affect strategic objectives, such as continuity of military operations or essential services, and can transform data into relevant tactical decisions [6].

In recent years, some technological CSA-devoted platforms have emerged to support the objectives of CSA. Military solutions, like the ones described in [4], [7], [8], aim to integrate operational data's cyber and physical perspectives in a hybrid decision-making system. In addition, different studies discuss the importance of visually transmitting data to non-expert personnel and analyze how this can be represented [9].

CSA platforms play a crucial role in military cyber defense by providing commanders with real-time information that can be interacted with. These intelligent platforms are responsible for transforming numerous data streams into concentrated and displayable information that allows officers to understand the operational context, enabling them to make tactical decisions that mitigate risks by anticipating potential risks and ensuring mission continuity. Thus, this paper aims to advance the comprehension of requirements and the need for a military smart CSA platform. This study seeks to find an open-source hybrid tool with cyber defense and asset management capabilities that can be used in both civilian and military environments.

Thus, the main contributions are summarized as follows:

- Study of 5 CSA platforms, identifying functional differences, integration capabilities, and architectures, identifying benefits and limitations of each proposal.
- Proposal of 6 criteria to be considered in designing a military smart CSA platform considering multi-domains: users, data sources, operational environments, supported use cases, visualization, and military requirements.
- Validation of the proposed criteria through the development of a cyber defense scenario deployed in an open-source platform for CSA (i.e., CRUSOE), allowing identification of gaps and opportunities for extension.

This paper is structured as follows: Section II reviews selected papers on existing CSA platforms, analyzing their main contributions. Section III compares different CSA platforms, identifying their main features. Section IV defines criteria required for a military smart CSA platform. Section V presents experiments conducted on CRUSOE. Finally, Section VI summarizes the conclusions and outlines future work.

II. State of the art

National defense organizations have long seen the need to address the cyber dimension as another field of operations to achieve dominance. The European Defense Agency (EDA) has highlighted, through its interest in its capability development programs funded by the European Defense Fund (EDF) [10], the importance of enhancing the cyber resilience of Member States and, more importantly, its specific interest in technological development projects focused around cyber situational awareness and cyber defence operations capabilities. Aligned with this call to increase SA scientific effort, this section details some of the most important works related to CSA.

As early as 2014, a Systematic Literature Review (SLR) on CSA was published [2], identifying CSA research thematic clusters and future research directions. Two key points of this study highlight, firstly, how most CSA research remains theoretical, as just a few studies validate real-world impact using empirical data. Secondly, the lack of a unified framework or toolset indicates that existing studies focus on specific applications but lack holistic integration.

One of the CSA research thematic clusters identified in this previous work is "Visualization". In line with this, Jiang et al. [7] extensively reviews different research works focusing on visualization to enhance CSA. They analyze different CSA proposals in terms of the maturity of visualizations and their relationship with key stakeholders, data sources, and CSA levels supported. This SLR notes some finding in the studied CSA works, such as: proposals primary revolve around technical users ($\sim 65\%$), there is a lack of projection capabilities in terms of Endsley's situational awareness theory [11] ($\sim 18\%$), and a rigorous and industry-suitable evaluation is not generally applied ($\sim 24\%$).

Deviating from SLRs, works such as of Georgi et al. [12] accurately stablishe guidelines and standards to follow in this research field, as it delves into metrics and methodologies used to assess SA, identifies key data sources recommended to enhance CSA, and proposes practical evaluation scenarios to measure the effectiveness of CSA techniques, e.g., APT and Insider Threat scenarios.

A solid foundation for SA innovation, but one that exhibits one of the previously identified gaps, is the work of Skopik et al. [9]. Such a paper introduces a Cyber Common Operating Picture (CCOP) framework and data model, designed to be flexible, scalable, and structured for the military domain. A case study tested for the Austrian Ministry of Defense (MoD) is also presented, demonstrating the feasibility of centralized cybersecurity reporting.

In contrast, the study by Ferreira et al. [8] demonstrates a strong commitment to military applications, based on a CSA architecture that integrates kinetic (physical) and cyber environments to enhance decision making and mission planning. Such a work prioritizes Mission Planning and Cyber Threat Mitigation (projection and action), yet this work shows that further implementation efforts are required instead.

Some of the studied proposals showcase the performance of tools with real development, as is the case of Llopis et al. [6], in which a comparison is made within the "Visualization" aspect of two actual military CSA tool developments (not public). The comparison is of great interest even though the identified challenges for both tools are equally important: lack of integration between cybersecurity and mission-oriented decision-making tools, high cognitive workload for users, and missing risk assessment models.

A notable paper bounded by the 2019 ECYSAP project (European Cyber Situational Awareness Platform) is the one presented by Sotelo et al. [13], which demonstrates the impact of Tactical Denial of Sustainability (ToDS) (negative effects on logistical or operative assets at the very edge of a military network due to cyberincidents).This proposal, which reinforces CSA platforms with advanced features, is also accompanied by other ECYSAP-associated work of Rodriguez et al. [14] where the effect of stress on CSA tool's operators and decision-makers is studied, recommending the incorporation of tools capable of accounting the effect of stress on the staff during cyber operations in CSA frameworks.

Previous works help identify some major gaps in the scientific effort surrounding SA. Scientific activity in this area is thematically diverse in an effort to explore this domain sufficiently to define a solid foundation for SA. However, this has led to a wide array of proposals of diverse thematic that, at present, remain largely unimplemented. Also, most existing studies focus primarily on the first two levels of the SA Endsley's model, consequently, there is a notable lack of proposals (and, more importantly, tools) with active assistance capabilities for projection-related tasks across all operational dimensions, e.g., mission planning, action execution, situation prediction, and impact analysis.

III. Functional comparison of CSA platforms

Technological CSA-devoted platforms improve threat detection and transform raw data from multiple sources into operationally relevant knowledge, which is essential for tactical, agile decisions. Thus, this section compares 5 CSA chosen platforms from a functional perspective. The comparative Table I summarizes the analysis results.

The comparison centers on 6 key functional aspects of CSA platforms. Below is an explanation of each of these aspects.

- **Users**: The target audience of the tool, i.e., decision makers with less technical knowledge, technical-operational personnel, or non-expert personnel.
- **Data sources**: The data sources received by the CSA platform, whether they come from its own devices or other entities, and whether it is specific defense or attack data.
- **Operational environments**: The proposed environments for deploying the CSA include Cybernetic Information Systems (CIS), which involve only electronic components or "passive entities", and Tactical Military settings, which encompass various military assets capable of active participation and response
- **Supported use cases**: When assessing which scenarios the different platforms can be applied to, the use cases

presented help to understand their flexibility and adaptability to different situations and operational needs.

- **Visualization techniques**: The transmission of information is a fundamental feature of CSA platforms, as effective transmission of information is crucial for fast and accurate decision making in complex environments. This criterion evaluates the different ways in which each platform presents and communicates critical information.
- **Military approach**: CSA platforms encompass a wide range of functionality applicable to many uses. This criterion focuses on extracting relevant functionalities that may be specifically useful in the military domain, allowing a more accurate assessment of their value in military strategies and planning and executing missions.

A. Proposal of Kookjin et al.

Kookjin et al. [4] underscored the importance of visualization and the human factor, proposing the design of a Cyber Common Operational Picture (CyCOP) that focuses on analyzing cyberspace operational nodes and aspects of UI. This work establishes a relationship between military mission planning, i.e., JCOPP (Cyberspace Operational Planning Process) & JTC (Joint Targeting Cycle), and the essential capabilities of an effective COP. However, their Proof of Concept (PoC) only implements stages for defining the operational environment and visualizing threats and assets in the hybrid space, leaving projection-related capabilities for future work.

B. Proposal of Esteve et al.

Esteve et al. [3] developed a hybrid model that georeferences cyber domain information from physical assets. Their CyCOP platform features an immersive visual representation of real-time elements and events, showing how threats affect multiple environmental components. The interface supports external data sources and allows manual user input through logical rules, though the specific implementation details are only named rather than explained.

C. Proposal of Noel et al.

Noel et al. [15] developed a PoC system for the US Army Command (C5ISR), enabling cyberspace visualization and understanding. The architecture features components for data transmission and integration of tools that build a graph knowledge base from incoming data streams, supporting multi-level data correlation. The system thoroughly implements Cyber Situational Understanding (Cyber SU) capabilities required for tactical mission support. However, as a PoC developed by the US Command, it is not publicly available.

D. Proposal of Dillabaugh et al.

Dillabaugh et al. [5] developed a PoC that enhances SA using a three-layer cyber-physical model. This system represents entities with nodes and employs simplified NATO Joint Military Symbology. The software, which was tested in defensive, offensive, and remediation scenarios, provides a detailed user guide and scenario explanations. However, it lacks flexibility for scenario modification and has limited data input options.

E. Proposal of Husak et al.

Husak et al. [16], [17] performed CSA in heterogeneous and complex environments related to network infrastructure. Seeking to cover the entire OODA loop (Observe, Orient, Decide, Act) through use cases, where we see as the only paper in which the decomposition of missions into objectives is developed through the capabilities related to the Decide phase, and where it is possible to interact with the various assets either to mitigate or obtain feedback through the functionality related to the Act.

F. Analysis of comparative table

After analyzing the 5 CSA platform proposals in the comparative Table I, an analysis has been carried out. In particular, most platforms target technical or technical-strategic teams, with Esteve's solution addressing all users across levels [3]. Data sources are predominantly internal, though some incorporate external or adversarial data. Although all platforms operate in cybernetic (CIS) environments, only Noel's solution extends to tactical military scenarios [15]. Common use cases include visualization of cyber knowledge graphs, mission impact assessment, and operational node visualization. All platforms prioritize visual representation through main maps or graph visualizations, with some offering advanced techniques like 3D modeling and virtual reality (Esteve). Military-specific features vary from NATO symbol integration (Dillabaugh) to mission dependency modeling (Noel) and decision support systems (Husak). Each platform demonstrates unique strengths: Kookjin in georeferencing and traffic visualization; Esteve offers immersive 3D and VR visualization; Noel provides comprehensive mission impact assessment; Dillabaugh emphasizes operational node visualization and mission engineering; while Husak focuses on incident handling and decision support. This comparison highlights the diverse approaches to CSA, with each platform tailoring its capabilities to specific operational needs and user groups.

The analyzed works show how CSA research focuses primarily on the first two levels of the SA process. For the civilian industry, available solutions such as Security Information and Event Management (SIEM, centered around aggregation and monitoring) are sufficient to improve the CSA level, since an industrial network is usually a homogeneous environment where only the cyber landscape is considered. More advanced solutions, such as SOAR, significantly enhance hardening capabilities and action execution within the network, aligning more closely with the "Act" category of the OODA loop.

Nevertheless, there is a notable lack of proposals or research on tools with active assistance capabilities for projection-related tasks, such as threat spread projection, mission decisive-condition identification, situation prediction, and impact analysis. This is the critical point for any NATO Federated Mission Networking service: having interoperability capabilities to aggregate information from C2ISR COI-services [18]

Proposal	Users	Data Sources	Operational environments	Supported use cases	Visualization techniques	Military requirements included
Kookjin [4]	Technical team	Own. Blue team	Cybernetic (CIS)	Georeferencing, cyber knowledge graph and traffic visualization	Small screens, Main map scorecard with controls and logs.	Visualization of operational nodes
Esteve [3]	Strategic Operational and Technical team	Own. Red and blue team	Cybernetic (CIS)	Hybrid situational, cyber knowledge graph and event classification	Main map , scorecard with 3D modeling and Virtual reality	Risk analysis, visualization of operational nodes
Noel [15]	Technical-strategic team	Own and external	Cybernetic (CIS), Tactical Military	CySU, mission impact assessment, threat detection	Main map, graph visualizations and status panel	Mission dependencies and functions, mission dependencies
Dillabaugh [5]	Cybernetic Emergency Response Team	Own. Red and blue team	Cybernetic(CIS)	Operational node visualization, Mission Engineering and cybernetic graph	Main map , Graphs with relations nodes, widget information	NATO symbols & mission symbols
Husak [16]	Technical team	Own	Cybernetic(CIS)	Incident handling decision support CySA	Cyber knowledge graph, mission enginering network defense management	Mission modeling fingerprinting decision support

Table I: Comparison of different CSA platforms

regardless of the domain (cyber or physical) and being able to infer information that assists decision-making profiles.

IV. Design of a military smart CSA platform

This section describes the 6 proposed criteria to be considered for the design of a military smart CSA platform, and then it introduces a high level architecture of platform built under such criteria.

We propose the concept of a military smart CSA platform to define a solution used in cyber defense operations, which leverages current market technologies to assist in data inference and decision-making. Such a platform should be able to provide a dynamic representation of the environment, both in terms of visual representation and managing the data plane, serving a wide array of user profiles and use cases, matching the following 3 main objectives:

- Multi-dimension: Operation over different domains, integrating both data collection and interpretation capabilities as well as the ability to generate new information for situation prediction and action execution over available assets across the different dimensions managed by the CSA platform.
- Relevant information according to user profile: Information selectively presented and gathered based on their relevance to different user profiles in a given situation within a specific time frame. This characteristic ensures a platform applicable to different scenarios and usable by technical, operational, and strategic profiles.
- Multi-domain: Aggregated information from diverse data sources, establishing relationships between the cyber and kinetic domains (land, sea, air, space), enabling hybridized outputs that facilitate technical-operational analysis and the projection of actions across the different dimensions.

A previous work of Georgi et al. [12] proposed quantitative evaluation criteria of CSA techniques and applications based on some aspects such as: timeliness, accuracy, and analyst's workload. Such mentioned aspects were obtained from the application of evaluation methodologies centered around the observation and interrogation of analysts, during and after a cyber-exercise, while using a CSA tool. However, in our case, we want to move away from a benchmark based on the tool's performance and aim to identify qualitative criteria related to capabilities that imply a cognitive or tactical advantage in a military scenario. Thus, the previous 3 main features associated with a military smart CSA platform spawn the following 6 criteria.

A. Criteria 1: Multi-users adaptability

Decision-making teams in civil cyber environments and hybrid military environments with tactical-strategic needs consist of diverse user profiles, e.g., tactical units, technicians, officers, and analysts. Thus, to maximize useful information delivery for each profile, a military smart CSA platform should adapt its interfaces and datasets to the user's expertise, avoiding the provision of irrelevant information.

B. Criteria 2: Multiple quality data sources

Most existing CSA platforms follow a data model that, at a minimum, fully covers cyber data ingestion and processing, e.g., topology analysis, identification of vulnerabilities in digital systems, and traffic analysis. Thus, a military smart CSA platform should extend their coverage to a broader range of quality data sources, e.g., data collection from SIEMs or SCADAs in industrial centers, interfaces with military C2 systems, and third-party datasets for analytical models. This variety of sources directly influences the tool's ability to operate across multiple physical dimensions by integrating data from diverse domains.

C. Criteria 3: Dual operational environments

CSA tools may operate in the cyber domain, showing certain hybrid capabilities such as the prediction of degradation of a

tactical asset due to a compromise in a managing digital node. This capability includes, for example, a situation where the surveillance capabilities of a drone system will be affected if the controlling swarm system suffers a performance problem. In addition, CSA tools may operate in the Mission domain, highly focused on supporting tactical-operational planning of military operations, having as main capabilities the visualization of operational nodes (facilities or units), the mission course analysis, and the analysis of communications effectiveness. Thus, a military smart CSA platform should support both Cybernetic (CIS) and Mission (MI) operational environments.

D. Criteria 4: CSA-level use cases

Minimal covered use cases involve data gathering derived from network and asset audit and its posterior visualization, e.g., network traffic monitoring and digital asset inventory. Such minimal covered use cases are aligned with the first 2 CSA levels: perception and comprehension. However, more innovative tools provide the third SA level (projection), e.g., decision support via situation prediction or high-value target identification. The projection level is strongly related to the final phase of the OODA loop, i.e., Act, which enables actions such as automatic configuration of firewalls or automatic shutdown of systems to avoid threat propagation. Thus, a military smart CSA platform should support use cases for the 3 CSA levels: perception, comprehension, and projection.

E. Criteria 5: Enriched visualization

Some works, e.g., [7], reveal limitations of CSA platforms in terms of visualization capabilities, which directly impact situational understanding. Even if it can be difficult to establish a full taxonomy of visualizations, it is possible to identify that some novel visualizations ease the comprehension of information and boost human inference. Thus, a military smart CSA platform should support minimal visualization techniques based on: knowledge graphs, C2-cybernetic hybridized views, and georeferenced views.

F. Criteria 6: Military approach

Finding references about CSA platforms for use in military environments may be difficult because of their confidentiality. However, some non-confidential works [3], [8] show that CSA platforms are widely used in defense. Thus, a military smart CSA platform should support different use cases that are quite specific for the defense sector and probably are not found in other sectors, e.g., visualization of operational nodes network and identification of high-value targets, compilation of a military-industrial COP, mission projection, and risk analysis on tactical assets, among others.

Throughout the study, numerous civilian and military architectural proposals have been examined. Still, the absence of a common architecture has led us to create one for a hybrid CSA platform applicable to both fields. Therefore, in Figure 1, it is possible to see the proposed design.

This design consists of 4 basic components identified as fundamental for a CSA platform. The first is responsible for collecting and sending information to the second module, where data aggregation occurs. After data correlation, the analysis module processes the information and generates a series of options that facilitate decision-making. Finally, this data is sent to the action module responsible for decision-making.

V. Experiments

This section develops the validation of the 6 proposed criteria for a military smart CSA platform. We will use a real open-source toolset as a testbed to simulate a cyber operation, observe its performance, and evaluate the degree to which each criterion is met, as well as how this affects overall CSA.

In the scientific literature used as a reference, we observe well-defined research trends, i.e., studies mainly focus on conceptually demonstrating the applicability of CSA in cyberwarfare and its short-term importance. However, a few proposals go beyond the theoretical framework and use current market technologies to shed light on how CSA tools impact real-world scenarios in operational-strategic decision-making.

Our experiment will consist of deploying and testing the most mature tool among the five proposals studied in Section III. The selected tool is developed from the work of Husák et al., the CRUSOE toolset. The reason for selecting CRUSOE is that it is a tool whose background, early development stages, final product (and later extensions) have been documented [19], [16], [17] over four years in various scientific publications. This tool was developed by the CSIRT-MU (Computer Security Incident Response Team of Masaryk University) and evaluated within the Masaryk University network for three months, creating a knowledge graph of 697,783 nodes and 22,119,299 edges with detailed information about 29,535 unique IP addresses in the network [16].

This toolset is open-source and finely crafted to be based around the OODA loop, featuring the following capabilities:

- Continuous data aggregation (active and passive network monitoring, third-party sourcing, local data sourcing).
- Visualization dashboard.
- Mission decomposition and mission resilience calculation.
- Search engine that identifies an asset's similar nodes in the knowledge graph.
- Threat spread projection graph.
- Integration with network security tools APIs to execute actions (firewall, DNS, mail filters, Active Directory, etc.).

The validation will be conducted considering a specific cyber defense scenario: an adaptation of the scenario described in the work of Dressler et al. [20], which considers a military logistical operation where kinetic-cybernetic cognitive capabilities must be deployed to fuse the threat surface dimensions into a single plane that allows identification of admissible mission trade-offs and success conditions. This hypothetical case is graphically represented in Figure 2. A U.S. Joint Task Force (JTF) heavily relies on a logistical-support information system designed to maintain personnel and supply logistics

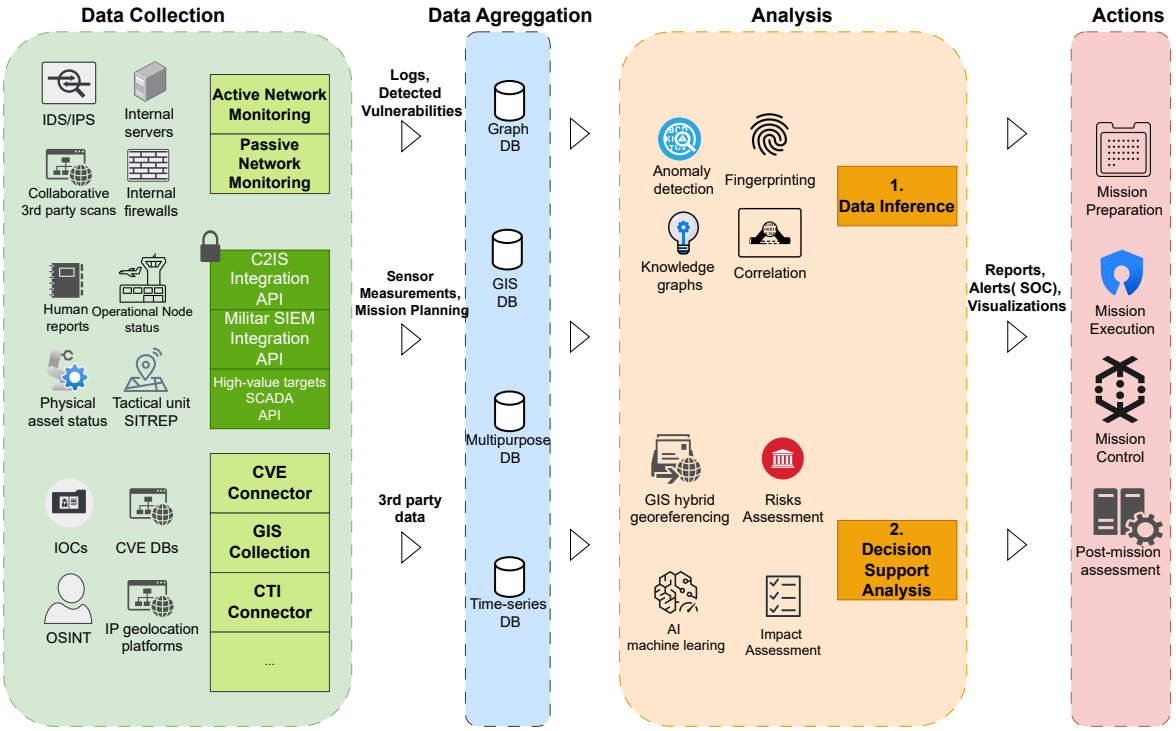


Figure 1: High level architecture of a Smart CSA platform

during combat operations, strategically defining this system as cyber key terrain. The supervision of these operations involves personnel with various roles: JTF commander, intelligence officers, and a cyber-forensics team.

The narrative of the use case presents allied forces with a series of conditions that deteriorate the defensive capability of the logistical network: intelligence reports indicating a change in the adversary's prioritized target list, and the discovery of a critical vulnerability in the logistical-support system. This scenario presents a set of variables that positively and negatively affect the capabilities of allied forces across different AOO dimensions and all phases of the OODA loop (Observe, Orient, Decide, Act): operational readiness of the cyber-forensics team, applicability of mitigation measures, intelligence reports provided by personnel, etc.

The experiments were conducted by deploying CRUSOE using Docker on a single server (4-core Intel N100 CPU, 8GB RAM), and the datasets used in the experiment were obtained from the official CRUSOE repository [21]. These datasets were loaded into CRUSOE's core database, Neo4j. Next, we describe the evaluation results of the six criteria proposed in Section IV for CRUSOE.

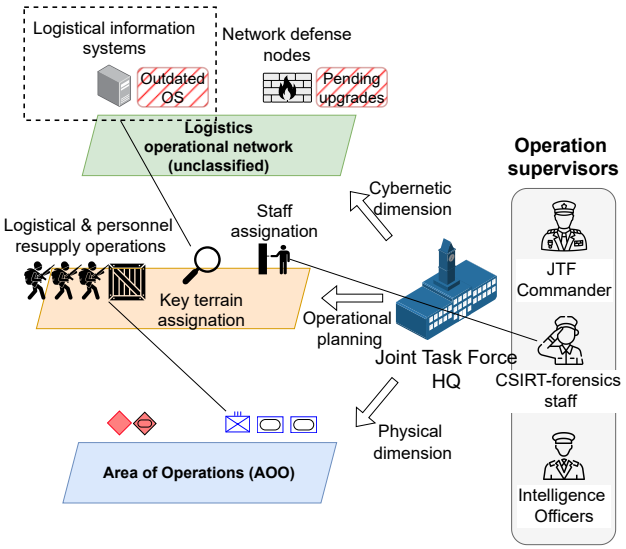


Figure 2: Graphic representation of Dressler [20] CSA use case

A. *Criteria 1: Multi-users adaptability*

CRUSOE's different modules achieve a certain level of profile adaptation. Its dashboard for monitoring all information is tailored for incident response teams that need an overview of the cybersecurity posture. Meanwhile, the Decision and Act functionalities have an interface designed for network operators involved in strategic planning but maintain an interactive format, easing the cognitive workload for technical experts and keeping a partially friendly UI for decision-makers.

CRUSOE partially meets the criteria for user adaptation in a smart CSA Platform. Its modular design effectively supports different user profiles but remains primarily framed within the

technical profile. There is no user-driven customization, such as dynamic interfaces that adapt based on expertise level or role-based information filtering.

B. Criteria 2: Multiple quality data sources

CRUSOE provides extensive cybersecurity monitoring and decision-support capabilities as it aggregates data from multiple cybersecurity sources, such as public vulnerability databases (e.g., NVD, OSVDB, Exploit Database) and manufacturer-specific vulnerability feeds (e.g., Cisco, Microsoft). Theoretically, it also combines passive network monitoring (via NetFlow/IPFIX) with active scanning (Nmap) to detect vulnerabilities in networked assets.

However, during the real use of CRUSOE, we observed limited active monitoring capabilities, as most data sourcing focused on the preloaded information. Indeed, a partial analysis of the real-world deployment was performed, as information from the physical host where the platform was deployed was detected and added to the Graph Representation DB, but no further scan capabilities were observed.

C. Criteria 3: Dual operational environments

The CRUSOE platform is primarily designed for cybersecurity situational awareness rather than tactical-operational military planning. Its architecture and functionalities focus on network security monitoring, vulnerability analysis, and automated response.

Although the Act capability of the CRUSOE is tightly coupled with security nodes, such as firewalls and switches, within the topology of the fictional scenario, there has been no real test of the performance of configuration profiles applied to network nodes due to the absence of a configurable network topology. This aligns with our statement about CRUSOE being more of a cyber-focused CSA tool than allowing mission planning and multi-domain operations. One example is the lack of kinetic dimension features, such as operator mission assignment, a commonly proposed capability in other works, and very useful for the selected use case.

D. Criteria 4: CSA-level use cases

The key element to choose CRUOSE as the starting target to benchmark is that it features at least one capability framed within each phase of the OODA loop. The scientific work and manuals [16], [17], [21] cite data aggregation based on key data sources (Observe), a well-fitted cyber data-model that allows the construction of knowledge graphs and ease the representation of data (Orient), a set of capabilities that allow "mission decomposition", "mission resilience calculation", "prediction of threat spreads", "data correlation optimized search-engine" (Decide) and finally multiple network services integrations with CRUSOE e.g., DNS, firewalls, mail- to allow for application of security profiles based on the infered information from the Decide phase's data.

As a matter of fact, our evaluation of this criterion using the toolset is not yet completed. Decision-assistance capabilities were not fully tested as the deployment of the complete toolset was not doable due to technical complications. Due to this same reason, only "missions decomposition" and "prediction of threat spread" have been tested regarding the Decision phase's capabilities. Still, the enabled subset of Act capabilities, along with the complete Observe capabilities, provides a cognitive value high enough to allow the users to visualize the cyber-environment and easily identify the attack surface of the high-value services identified by the Network Scan Data, enhancing the overall level of CSA of the operator.

E. Criteria 5: Enriched visualization

CRUSOE utilizes network graphs, dashboards, and attack modeling to help analysts identify security risks. However, its visualization capabilities are primarily limited to cyber network representations and do not fully extend into hybridized views that integrate cyber and kinetic domains.

Most visualization work is done through CRUSOE's dashboard (Orient module). While CRUSOE effectively represents cyber data, it lacks more advanced visualization models such as geospatial overlays, heatmaps, or interactive hybrid dashboards. Its primarily graph-oriented dashboard visualization is also not easily navigable, as in some cases, access to the graph-representation DB is necessary to study omitted needed to better understand the context of the knowledge graph presented in the main dashboard.

F. Criteria 6: Military approach

CRUSOE provides strong capabilities in network monitoring, risk assessment, and automated response. However, its alignment with military-specific use cases appears to be limited based on the documentation and conducted benchmark. On paper, it features different APIs that allow CRUSOE to interact with external services such as firewall configurations, DNS, and mail filters, suggesting the possibility of developing APIs to integrate SCADA or C2 systems.

Currently, the most attractive feature for military cyber defense is the Decide capabilities. This capability uses attack graphs and Bayesian inference to predict security breaches, a method that could potentially be adapted to military risk analysis on tactical assets.

VI. Conclusions and Future Work

In this paper, we have compared different CSA platforms. This study has been done through the search of different proposals and studies carried out, where different architectures are proposed, what functionality they offer, and whether they are conceptual studies or if the tool has been developed. To make this comparison, a table was created considering a series of proposed criteria. These criteria have been presented to support future developments and used to evaluate the different studies. Finally, the CRUSOE platform [22] has been used to test the criteria defined in this tool to measure its degree of maturity and identify expansion opportunities.

CRUSOE has several limitations when used in a military-grade CSA platform, due it is designed primarily for civilian network management. It lacks hierarchical data and decision

processing, military classification levels for chain-of-command protocols, and military-oriented features in the ACT component, including geospatial visualization and tools for targeted attack analysis. Additionally, its reliance on Neo4j results in high costs for military use, and it is not equipped for multi-domain or multinational joint operations due to scalability issues. To improve CRUSOE's suitability for military applications, enhancements are needed in the Decide and Act phases, which currently demand the most research. This involves expanding its capabilities to operate in dual-operation enviroment by integrating tools to manage complex assets with dependencies across various dimensions.

Future work will explore the specialization of the CRUSOE platform as a CSA tool that can be effectively employed by military commanders. Additionally, we have proposed a set of qualitative criteria to evaluate and develop SA tools, enabling assessment of the tactical and cognitive advantages these tools provide in military scenarios.

Acknowledgment

This work has been co-funded by the European Union (project ECYSAP EYE). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Defence Fund. Neither the European Union nor the granting authority can be held responsible for them. This work has also been partially supported by MCIN/AEI/10.13039/501100011033 NextGeneration EU/PRTR, UE, under Grant TED2021-129300B-I00, by MCIN/AEI/10.13039/501100011033/FEDER, UE, under Grant PID2021-122466OB-I00, by the Spanish National Institute of Cybersecurity (INCIBE) by the Recovery, Transformation and Resilience Plan, Next Generation EU under the strategic project DEFENDER, by the CyberDataLab (Cybersecurity and Data Science Laboratory) at the University of Murcia (Spain), and the School of Engineering, Science and Technology at the University of Rosario (Colombia).

References

[1] T. F. Ask, B. J. Knox, R. G. Lugo, L. Hoffmann, and S. Sütterlin, "Gamification as a neuroergonomic approach to improving interpersonal situational awareness in cyber defense," *Frontiers in Education*, vol. 8, 2023. [Online]. Available: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.988043

[2] U. Franke and J. Brynielsson, "Cyber situational awareness – a systematic review of the literature," *Computers & Security*, vol. 46, pp. 18–31, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404814001011

[3] M. Esteve, I. Pérez, C. Palau, F. Carvajal, J. Hingant, M. A. Fresneda, and J. P. Sierra, "Cyber common operational picture: A tool for cyber hybrid situational awareness improvement," *North Atlantic Treaty Organization (NATO) Science and Technology Organization (STO), Technical Report STO-MP-IST-148*, 2016.

[4] K. Kim, J. Youn, S. Yoon, J. Kang, K. Kim, and D. Shin, "Study on cyber common operational picture framework for cyber situational awareness," *Applied Sciences*, vol. 13, no. 4, p. 2331, 2023.

[5] C. Dillabaugh and D. Bennett, "Cybercop: Cyber situational awareness demonstration tool," *Defence Research and Development Canada, Ottawa, Canada*, 2020.

[6] S. Llopis, J. Hingant, I. Pérez, M. Esteve, F. Carvajal, W. Mees, and T. Debatty, "A comparative analysis of visualisation techniques to achieve cyber situational awareness in the military," in *2018 International Conference on Military Communications and Information Systems (ICMCIS)*, 2018, pp. 1–7.

[7] L. Jiang, A. Jayatilaka, M. Nasim, M. Grobler, M. Zahedi, and M. Ali Babar, "Systematic literature review on cyber situational awareness visualizations," *IEEE Access*, vol. 10, pp. 1–1, 01 2022.

[8] A. F. A. Machado and E. T. Yano, "Conceptual architecture for obtaining cyber situational awareness," 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:108175837

[9] F. Skopik, A. Bonitz, V. Grantz, and G. Göhler, "From scattered data to actionable knowledge: flexible cyber security reporting in the military domain," *International Journal of Information Security*, vol. 21, pp. 1–25, 09 2022.

[10] "European Defense Fund work programme 2025," https://defence-industry-space.ec.europa.eu/edf-work-programme-2025_en, accessed: 2025-03-20.

[11] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995. [Online]. Available: https://doi.org/10.1518/001872095779049543

[12] G. Nikolov, A. Perez, and W. Mees, "Evaluation of cyber situation awareness - theory, techniques and applications," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ser. ARES '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3664476.3670921

[13] M. A. Sotelo Monge and J. Maestre Vidal, "Conceptualization and cases of study on cyber operations against the sustainability of the tactical edge," *Future Generation Computer Systems*, vol. 125, pp. 869–890, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21002788

[14] D. Sandoval Rodríguez-Bermejo, J. Maestre Vidal, and J. Estévez Tapiador, "The stress as adversarial factor for cyber decision making," in *Proceedings of the 16th International Conference on Availability, Reliability and Security*, ser. ARES '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3465481.3470047

[15] S. Noel, S. Purdy, A. O'Rourke, E. Overly, B. Chen, C. DiFonzo, J. Chen, G. Sakellis, M. Hegde, M. Sapra, C. Araki, J. Martin, B. Koehler, J. Keenan, T. Coen, W. W. Watson, J. Harper, and K. Jacobs, "Graph analytics and visualization for cyber situational understanding," *The Journal of Defense Modeling and Simulation*, vol. 20, no. 1, pp. 81–95, 2023. [Online]. Available: https://doi.org/10.1177/15485129211051385

[16] M. Husák, L. Sadlek, S. Špaček, M. Laštovička, M. Javorník, and J. Komárková, "Crusoe: A toolset for cyber situational awareness and decision support in incident handling," *Computers & Security*, vol. 115, p. 102609, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404822000086

[17] M. Husák, L. Sadlek, M. Hesko, V. Šebela, and S. Špaček, "The evolution of the crusoe toolset: Enhancing decision support in network security management," in *2024 20th International Conference on Network and Service Management (CNSM)*, 2024, pp. 1–3.

[18] NATO Communications and Information Agency, "C3 taxonomy baseline 6," NATO C3 Board, Brussels, Belgium, Technical Report, October 2023, available at: https://www.nato.int/cps/en/natohq/topics_157573.htm.

[19] M. Husák, M. Laštovička, and D. Tovarňák, "System for continuous collection of contextual information for network security management and incident handling," in *Proceedings of the 16th International Conference on Availability, Reliability and Security*, ser. ARES '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3465481.3470037

[20] J. Dressler, C. L. Bowen, W. Moody, and J. Koepke, "Operational data classes for establishing situational awareness in cyberspace," in *2014 6th International Conference On Cyber Conflict (CyCon 2014)*, 2014, pp. 175–186.

[21] "original CSIRT-MU CRUSOE, howpublished = https://github.com/csirt-mu/crusoe, note = Accessed: 2025-03-21."

[22] "UMU CRUSOE testing environment, howpublished = https://github.com/anthonyfa/cdlcrusoe/tree/dev, note = Accessed: 2025-03-21."

Ciberseguridad Marítima en la era digital: Evaluación y validación de la planificación en POSEIDON

Ferney MARTÍNEZ
GSYA Research Group – Grupo PRODIN
Universidad Castilla-La Mancha – COTECMAR
Cartagena, Colombia
Ferney.Martinez@alu.uclm.es

José Luis RUIZ CATALÁN
GSYA Research Group – GSyA
Universidad Castilla La Mancha
Ciudad Real, España
joseluis.ruiz@uclm.ess

Luis Enrique SÁNCHEZ
GSYA Research Group – GSyA
Universidad Castilla La Mancha
Ciudad Real, España
Luise.Sanchez@uclm.es

Antonio SANTOS-OLMO
GSYA Research Group – GSyA
Universidad Castilla La Mancha
Ciudad Real, España
Antonio.Santosolmo@uclm.es

David G. ROSADO
GSYA Research Group – GSyA
Universidad Castilla La Mancha
Ciudad Real, España
David.GRosado@uclm.es

Eduardo FERNÁNDEZ-MEDINA
GSYA Research Group – GSyA
Universidad Castilla La Mancha
Ciudad Real, España
Eduardo.Fdezmedina@uclm.es

Resumen- **El sector marítimo es crucial para la economía global, pero su digitalización lo expone a crecientes amenazas cibernéticas. POSEIDON, un marco de ciberseguridad diseñado para la industria marítima fortalece la resiliencia mediante la gestión de riesgos, planificación estratégica y respuesta a incidentes. Este estudio evalúa la fase de planificación, una de las seis fases del marco de trabajo, validándola en un buque de propósito dual que sufrió un ciberataque. La aplicación de este enfoque estructurado evidenció mejoras en detección de amenazas, reducción del tiempo de respuesta y capacidad de recuperación. Estos hallazgos resaltan la importancia de una planificación proactiva, supervisión continua, formación del personal y protocolos estrictos para proveedores. Este estudio establece bases sólidas para aplicar POSEIDON en otras infraestructuras marítimas, reforzando la necesidad de un marco adaptable y dinámico ante los desafíos emergentes en ciberseguridad marítima.**

***Index Terms*- Ciberseguridad Marco POSEIDON, Gestión de Riesgos, Seguridad Marítima.**

Tipo de contribución: *Investigación original*

I. Introducción

El sector marítimo es un importante pilar para el mercado global que asegura el movimiento de cerca de un 90% del transporte de recursos comerciales y vitales a nivel mundial [1]. Cuenta con un amplio ecosistema que incluye embarcaciones comerciales, instalaciones portuarias, infraestructuras críticas y sistemas tecnológicos avanzados que permiten una navegación y comunicación eficientes. Sin embargo, este sector enfrenta la creciente amenaza de la ciberdelincuencia que día a día aumenta exponencialmente en factores de frecuencia y sofisticación [2-4], afectando sistemas vitales para la operatividad del buque como la navegación, las comunicaciones y la gestión portuaria. Esta situación impacta de manera directa causando graves interrupciones en la cadena de suministro global, llegando a comprometer datos sensibles y poner en riesgo la seguridad de tripulaciones y operaciones.

La digitalización acelerada del sector marítimo ha traído consigo avances técnicos y operativos significativos. No obstante, también ha incrementado la exposición a riesgos cibernéticos, lo que exige estrategias de protección más robustas. Los principales referentes de ciberseguridad que se aplican en la industria marítima se comparan en la Tabla I, que destaca sus enfoques metodológicos, casos de uso y oportunidades de integración en entornos navales.

Los sistemas interconectados, como el Sistema de Identificación Automática (AIS), el Sistema de Información y Visualización Cartográfica Electrónica (ECDIS), los Sistemas de Posicionamiento Global (GPS) y los sistemas de los Sistemas de Gestión de Tráfico Marítimo (VTS), son objetivos comunes de ciberataques por su importancia en las funciones básicas de operatividad [5,6]. Desde ataques que buscan paralizar la actividad de los puertos o buques mediante ransomware hasta la manipulación de datos de radar, estos incidentes evidencian las deficiencias en infraestructuras críticas, resaltando la necesidad de mejorar su nivel de preparación para enfrentar estas amenazas.

El marco de trabajo POSEIDON permite la gestión total de los riesgos cibernéticos para asegurar la continuidad operativa y el cumplimiento de la normativa internacional. POSEIDON presenta una metodología rigurosa que abarca todas las fases de la seguridad. Además, sus módulos permiten abordar desde situaciones cotidianas hasta crisis complejas, garantizando respuestas acordes a cada tipo de incidente.

El presente artículo destaca diversos aspectos del marco de trabajo POSEIDON, incluyendo sus fases, su estructura y las interrelaciones entre ellas, elementos fundamentales para fortalecer la resiliencia ante el actual escenario de amenazas.

Asimismo, el artículo destaca la importancia del trabajo conjunto entre los diferentes actores del sector marítimo. La implementación efectiva de POSEIDON requiere la adopción de estándares internacionales y la participación de operadores de embarcaciones, administradores portuarios, reguladores y expertos en ciberseguridad. En concreto, el estudio detalla la evaluación de la fase de la planificación y su implementación en un caso práctico.

En la sección II se presentan los antecedentes, donde se revisan los principales marcos regulatorios y metodologías aplicadas a la gestión de riesgos en el sector marítimo. A continuación, la sección III se describen los bloques y componentes de todo el marco de trabajo POSEIDON, detallando sus principios fundamentales y su estructura modular, en la sección IV se desarrolla focalmente el bloque de etapas con especial énfasis en la fase de planificación.

TABLA I
REFERENTES INTERNACIONALES EN CIBERSEGURIDAD APLICADOS AL SECTOR MARÍTIMO

Marco de Ciberseguridad	Origen	Objetivo y Aplicabilidad	Enfoque Metodológico	Ámbito de Uso	Certificación
NIST Cybersecurity Framework (CSF)	Instituto Nacional de Estándares y Tecnología (EE. UU.)	Optimizar la gestión de riesgos organizacionales mediante las funciones: identificar, proteger, detectar, responder y recuperar. Aplicable a infraestructuras críticas, incluyendo la marítima.	Funcional: Identificar, Proteger, Detectar, Responder, Recuperar	Organizaciones públicas y privadas, infraestructuras críticas	No
ISO/IEC 27000 Series (incl. ISO 27001, IEC 62443, IEC 61162-460)	Organización Internacional de Normalización (ISO)	Establecer e implementar un Sistema de Gestión de Seguridad de la Información (SGSI). Integra marcos complementarios aplicables al entorno naval.	Gestión: Plan-Do-Check-Act (PDCA)	Empresas, gobiernos, operadores marítimos	Sí
MITRE ATT&CK	MITRE Corporation (EE. UU.)	Taxonomía detallada de tácticas, técnicas y procedimientos (TTPs) utilizados por actores maliciosos. Base para detección avanzada de amenazas persistentes.	Táctico-operacional: TTP-basado en matrices	SOCs, equipos de respuesta, industria defensiva	No (formaciones externas)
COBIT 5 / COBIT 2019	ISACA	Gobernanza y gestión del riesgo TIC. Complementa marcos técnicos con una perspectiva organizacional.	Gobernanza, Planificar, Construir, Ejecutar, Monitorear	Entornos corporativos y tecnológicos complejos	Sí
BSI IT-Grundschutz	Oficina Federal de Seguridad de la Información (Alemania)	Proporciona un enfoque sistemático y práctico para implementar SGSI. Uso mayoritario en Europa. Compatible con ISO/IEC 27001.	Modular: Catálogo de controles por áreas temáticas	Administraciones públicas y sectores críticos europeos	Sí
OWASP (Open Web Application Security Project)	OWASP Foundation	Recursos abiertos para mejorar la seguridad en aplicaciones web y APIs. Relevante para plataformas digitales portuarias e interfaces embarcadas.	Colaborativo: Open source / Top 10 / MASVS	Desarrolladores, integradores, sector educativo	Parcial (formación no certificada formalmente)

Posteriormente, la sección V se expone el caso de estudio componente de planificación, evaluando su impacto en la resiliencia cibernética. En esta sección, se analizan los resultados obtenidos a través de métricas cuantificables que demuestran la efectividad de las estrategias implementadas. Finalmente, en la sección VI se registran las conclusiones discutiendo los hallazgos más relevantes, resaltando la importancia de una planificación proactiva en la ciberseguridad marítima y proponiendo futuras investigaciones para ampliar la validación del marco en distintos entornos operativos.

II. ANTECEDENTES

Por medio de la revisión sistemática de la literatura realizada en [7] se realiza un estudio a los principales marcos regulatorios y metodologías aplicadas a la gestión de riesgos en el ámbito marítimo. Se identifican amenazas clave que implican la manipulación de AIS [8,9], GPS [10, 11] y ECDIS [12], cruciales para la seguridad marítima y la navegación segura.

La investigación propone un marco de ciberseguridad al que llama POSEIDON, y plantea un conjunto de fases y procesos adaptados al entorno marítimo, todo diseñado para fortalecer la planeación, detección, prevención y respuesta ante amenazas cibernéticas. Este framework utiliza una metodología que fusiona la alineación estratégica, las medidas de seguridad preventivas, la resiliencia operativa y como pilar estratégico la gestión del talento humano y la generación de una cultura de la ciberseguridad para abordar los problemas de protección digital en las operaciones marítimas.

Esta investigación destaca la importancia de una estrategia multidisciplinaria, fusionar la ingeniería, la supervisión de riesgos y las reglas globales para reducir las debilidades y mejorar la durabilidad de los marcos marítimos contra los ataques cibernéticos.

De igual manera, durante las IX Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2024) se presentó un avance académico centrado en el desarrollo de un patrón de análisis de riesgos, como parte del proceso de evolución del Marco de Ciberseguridad POSEIDON -Procedimientos, Operaciones y Estándares para la Detección de Eventos Integrales y el Desarrollo de la Capacidad Operativa Naval-. La herramienta de implementación web llamada eMARISMA [13], proporciona un enfoque sistemático para la identificación, evaluación y mitigación de riesgos cibernéticos en infraestructuras marítimas en tiempo real. La propuesta es una respuesta a las necesidades del sector y se alinea con marcos regulatorios internacionales como ISO [14, 15], ENISA [16, 17] y NIST [18]. La investigación, apoyada por un astillero latinoamericano, incluyó una validación práctica en un buque de propósito dual. Esto permitió evidenciar la capacidad del modelo para fortalecer la protección ante vulnerabilidades y mejorar la resiliencia cibernética en entornos marítimos reales.

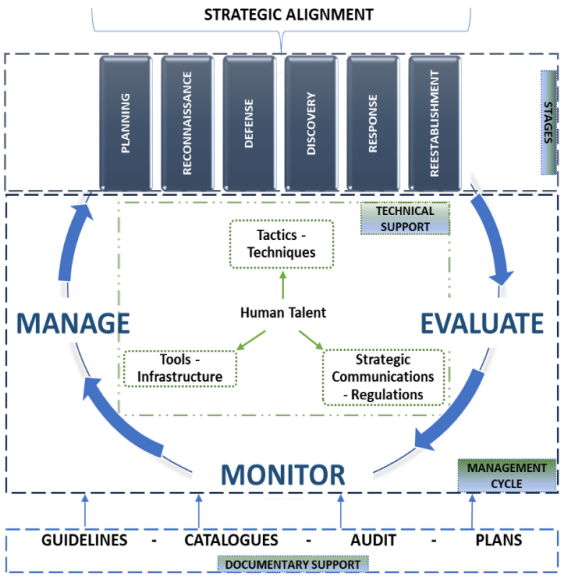


Fig. 1. Bloques y Componentes de POSEIDON

Los resultados socializados evidenciaron que MARISMA-SHIP, lograba integrar las estrategias de seguridad de aporte al sector, y aportar significativamente en la capacidad de respuesta ante eventos cibernéticos.

A partir de lo anterior, en el siguiente capítulo se presentará el primer nivel de las fases del marco de trabajo. Posteriormente, se profundizará en la fase de planificación dentro del marco POSEIDON, presentado a alto nivel en la Fig. 1, detallando sus procesos, subprocesos y sus interrelaciones.

III. BLOQUES Y COMPONENTES DE POSEIDON

POSEIDON incorpora procesos adaptativos con un enfoque modular, permitiendo que pueda darse de una manera natural la protección de activos críticos mediante la identificación de riesgos, la mitigación de amenazas y la recuperación eficiente tras incidentes. Cuenta con una estructura flexible con el objetivo que pueda ser aplicada en diferentes contextos como escenarios comerciales y militares.

La relevancia de establecer un marco de ciberseguridad como POSEIDON radica en su habilidad para examinar, prever y tratar los riesgos que surgen desde un ambiente marítimo en constante cambio. Esta sección presenta los principios y las fases del marco, enmarcando los procesos de la fase de Planificación que permiten su aplicación efectiva y su alineación con las necesidades operativas del sector.

La efectividad de POSEIDON se basa en su enfoque organizado pero versátil, formado por tres principios esenciales:

- Alineación Estratégica: Los procesos deben alinearse con los objetivos marítimos de alto nivel, como la continuidad operativa, la protección de infraestructuras críticas y el cumplimiento de normativas internacionales.
- Modularidad por Componentes: Cada proceso se construye teniendo como base elementos fundamentales como la Planificación, Reconocimiento, Defensa, Descubrimiento, Respuesta y Restablecimiento, garantizando su capacidad de adaptación a diferentes contextos marítimos.
- Mejora Continua: Los procesos se evalúan constantemente y son optimizados por medio de ciclos de retroalimentación que funcionan de manera iterativa.

En el siguiente capítulo, se abordará específicamente el bloque de etapas del marco de trabajo, detallando cada una de sus fases y su contribución a la resiliencia cibernética en entornos marítimos. Además, se iniciará el análisis específico del componente de planificación, explorando su estructura, sus interacciones con otros procesos del marco y su impacto en la gestión proactiva de la ciberseguridad marítima.

IV. BLOQUE DE ETAPAS EN EL MARCO POSEIDON

El marco de ciberseguridad POSEIDON adopta un enfoque sistémico para lograr realizar una gestión del riesgo cibernético en el sector marítimo. Su estructura secuencial: Planificar, Reconocer, Defender, Descubrir, Responder y Restablecer, busca adelantar procesos que permitan anticipar y mitigar amenazas en infraestructuras críticas, en especial la marítima. Este modelo resalta la importancia de la Planificación al incluirla como una fase principal, diferenciándola de otros enfoques centrados en la respuesta a incidentes. POSEIDON en el bloque de Etapas se subdivide en 6 componentes clave:

- Planificar: Contempla desde una etapa inicial la evaluación de riesgos, la asignación de recursos y el establecimiento de políticas y procedimientos ya que se alimenta de los lineamientos estratégicos. Considera factores específicos del sector naval, como perfil del personal y tipo de operación y las lecciones aprendidas para mejorar la gestión de la ciberseguridad.

Tabla II
RELACIÓN SUBCOMPONENTES POSEIDON CON PLANIFICACIÓN

	Subcomponentes de Planificación					
	Prevención	**Perfil del Personal**	**Lecciones Aprendidas**	**Prevención Operacional**	**Medidas con Terceros**	**Gobernanza de Seguridad**
Prevención	-	E	E	E	S	-
Perfil del Personal	S	-	S	S	E	E
Lecciones Aprendidas	E	S	-	S	E	E
Prevención Operacional	-	E	S	-	-	-
Medidas con Terceros	S	-	S	E	-	E
Gobernanza de Seguridad	E	-	E	E	-	-
Gestión de Actividades	E	-	E	-	-	-
Evaluación de Riesgos de Gobernanza	E	-	E	S	S	-
Estrategia de Gestión de Riesgos	E	S	-	E	S	-
Clasificación de Datos y Misiones	-	-	S	E	-	S
Registro de Incidentes	E	E	S	-	-	E
Gestión de Identidad	E	E	-	E	-	-
Seguridad de Datos	E	E	E	E	-	S
Protección de Información	E	S	S	S	-	-
Procesos y Procedimientos	E	S	-	-	E	-
Mantenimiento	-	E	E	E	E	-
Tecnología de Protección	E	-	E	-	-	E
Cultura de Seguridad Marítima	S	S	E	-	-	S
Monitoreo de Seguridad Marítima	S	E	E	E	E	-
Protección de Medios Digitales	S	S	-	-	E	S
Anomalías y Eventos	E	E	-	S	-	E
Monitoreo Continuo de Seguridad	S	-	-	-	E	S
Detección de Procesos	E	E	-	-	E	-
Gestión de Escenarios	E	-	-	S	S	-
Planificación de Respuestas	E	-	E	E	-	S
Comunicaciones	S	E	-	-	-	E
Análisis	E	E	S	S	E	-
Mitigación	S	-	-	E	-	-
Mejoras en Recuperación	E	-	-	E	E	E
Capas de Defensa en Profundidad	S	-	S	S	S	S
Planificación de Recuperación	E	-	E	-	-	-
Mejoras en Restablecimiento	E	E	E	E	-	-
Comunicaciones	S	S	S	S	-	-
Redes Dedicadas	-	E	-	E	-	-
Consolidado:						
TOTAL SALIDAS	10	7	9	9	5	7
TOTAL ENTRADAS	19	13	12	14	10	8
TOTAL SIN RELACIÓN	5	14	13	11	19	19

- Reconocer: Se enfoca en identificar activos, vulnerabilidades y amenazas en el entorno marítimo. Emplea atributos específicos como la clasificación por misión para una mejor gestión del riesgo. Considera elementos internos y externos operacionales como sistemas tecnológicos y perfil de la tripulación del buque.
- Defender: Aplica controles y acciones para la salvaguarda de sistemas y bienes marítimos. Incorpora tácticas de seguridad de múltiples niveles, resguardando sensores, actuadores y dispositivos esenciales. Fomenta un aspecto fundamental en este modelo como la educación en cultura de seguridad, incentivando buenas prácticas y la observancia de las regulaciones.
- Descubrir: Desarrolla capacidades de detección temprana de amenazas en el sector marítimo. Emplea técnicas y herramientas para modelar escenarios navales y analizar tácticas adversarias. Considera vulnerabilidades externas y riesgos en la cadena de suministro para mejorar la vigilancia.
- Responder: Organiza medidas para atenuar ataques y manejar incidentes a través de planes de respuesta organizados. Incluye estudios forenses y recolección de pruebas para robustecer la defensa cibernética. Promueve estrategias de cultura de la ciberseguridad como herramienta eficaz frente a los ciberataques.
- Restablecer: Elabora tácticas y aplica todos los procesos para la restauración de servicios después de un suceso cibernético. De igual manera, proporciona prioridad a la recuperación operativa de acuerdo con la misión y el efecto en la infraestructura del mar.

En este artículo, el foco principal se centrará en el componente de Planificación, debido a su naturaleza innovadora en la propuesta del marco POSEIDON y se descompone en la Fig. 2.

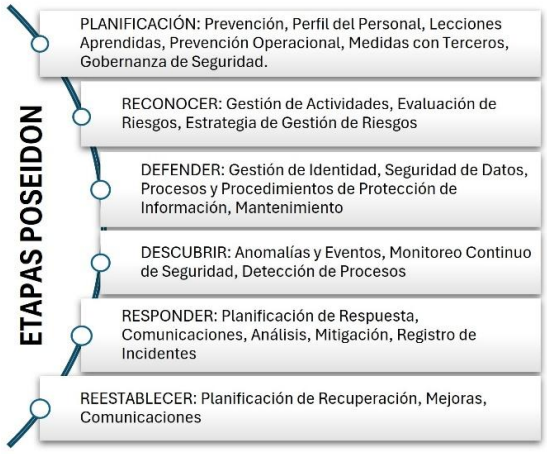


Fig. 2. Aportes del bloque de Etapas de POSEIDON

A diferencia de los enfoques o esquemas tradicionales que valoran la detección y respuesta a incidentes, la planificación promueve una gestión proactiva del riesgo, estableciendo estrategias, objetivos y controles antes de que las amenazas se transformen en realidad, lo que la convierte en una propuesta proactiva. Este componente es crucial para incrementar la capacidad de resistencia del sector, adaptando las operaciones a las normativas internacionales y asegurando una repartición eficiente de recursos. Además, su inclusión de aspectos específicos del sector naval, como las lecciones aprendidas, el perfil del personal y el método de operación, lo convierten en un elemento único que potencia la seguridad en entornos marítimos.

A. *Análisis de los subcomponentes de Planificación*

La Tabla II consolida un análisis de las relaciones entre los subcomponentes del marco de ciberseguridad POSEIDON y los del componente de planificación. Esta relación se establece evaluando si cada subcomponente en la fila actúa como entrada (E) o salida (S) dentro del proceso de planificación. A partir de estos resultados, se analiza su impacto en la implementación de POSEIDON.

A continuación, se examinan los subcomponentes del proceso de planificación: Prevención, Perfil del Personal, Lecciones Aprendidas, Prevención Operacional, Medidas con Terceros y Gobernanza de Seguridad, destacando su papel en la protección de activos críticos, la formación de la tripulación y la supervisión de estándares de seguridad en la cadena de suministro. Este análisis proporciona una visión integral de cómo estos elementos contribuyen a la construcción de un ecosistema marítimo más seguro, permitiendo una mejor anticipación y respuesta ante ciberataques y otros riesgos operativos:

Prevención

Como se visualiza en consolidación de la Tabla II, el subcomponente de Prevención tiene 19 entradas y 10 salidas, lo que establece la importancia en el recibo de gran cantidad de información para reforzar la seguridad en el sector marítimo. Este subcomponente depende de datos provenientes de otras áreas como Lecciones Aprendidas y Registro de Incidentes, lo que le permite mejorar continuamente las estrategias de seguridad. Su influencia en la capacitación del Perfil del Personal y en la implementación de medidas para terceros demuestra su impacto en la formación de la tripulación y en la extensión de estrategias de seguridad a toda la cadena de suministro.

Perfil del Personal

Con 13 entradas y 7 salidas, el Perfil del Personal refleja su papel clave en la ciberseguridad marítima. Este subcomponente recibe información de Prevención y Seguridad de Datos, lo que le permite establecer criterios de formación para la tripulación y el personal portuario.

La conexión que tiene con Prevención Operacional y Monitoreo de Seguridad Marítima indica su importancia en la implementación de medidas de ciberseguridad en todas las operaciones diarias.

Lecciones Aprendidas

El subcomponente de Lecciones Aprendidas muestra 12 entradas y 9 salidas, mostrando un equilibrio en el flujo de información. Recibe datos de Registro de Incidentes y Evaluación de Riesgos de Gobernanza, lo que permite optimizar estrategias de seguridad. Su salida hacia Prevención Operacional indica que las experiencias previas se utilizan para mejorar protocolos de seguridad y evitar la repetición de errores. Su influencia en la Cultura de Seguridad Marítima establece como el aprendizaje continuo de las tripulaciones fortalece la conciencia y la preparación ante amenazas.

Prevención Operacional

Con 14 entradas y 9 salidas, es un subcomponente clave en la ejecución de medidas de seguridad. Obtiene información de Prevención y Monitoreo de Seguridad Marítima, lo que le permite actuar en tiempo real frente a amenazas emergentes y asegurar que sea dinámica y adaptable.

Medidas con Terceros

El subcomponente de Medidas con Terceros tiene 10 entradas y 5 salidas, lo que indica que recibe información de diversas fuentes para garantizar que los proveedores y socios cumplan con los estándares de seguridad establecidos

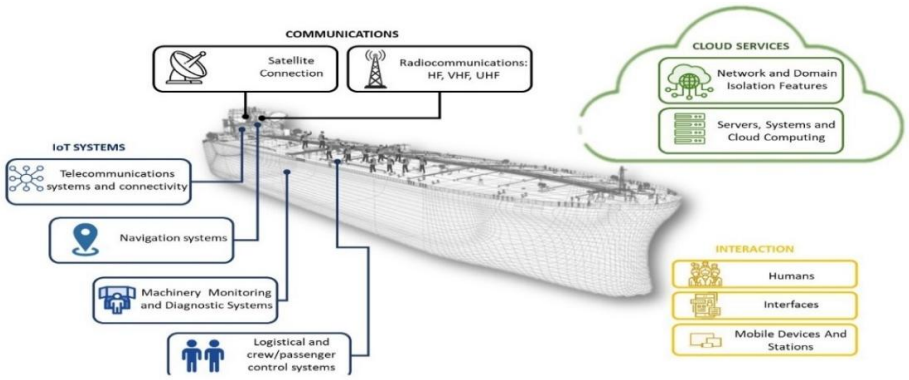


Fig. 3. Sistemas a bordo de buque dual

Recibe datos de Prevención y Evaluación de Riesgos de Gobernanza, lo que permite asegurar que los proveedores implementen

medidas adecuadas de ciberseguridad. La menor cantidad de salidas sugiere que, aunque es un elemento crítico para la seguridad en la cadena de suministro, su influencia directa en otros subcomponentes es más limitada.

Gobernanza de Seguridad

Este subcomponente presenta 8 entradas y 7 salidas, lo que refleja su papel como regulador de la seguridad en el sector marítimo. Recibe información de Tecnología de Protección y Anomalías y Eventos, asegurando que todas las medidas implementadas en los buques cumplan con normativas internacionales como las de la OMI. Sus salidas hacia Seguridad de Datos y Protección de Medios Digitales evidencian que la gobernanza no solo dicta pautas internas, sino que también impacta en la protección de los socios comerciales y proveedores.

El análisis de los subcomponentes revela que la prevención es el núcleo principal del sistema de planificación de seguridad, con la mayor cantidad de ingresos y egresos. El Perfil del Personal y las Lecciones Aprendidas juegan un papel vital en la capacitación y el perfeccionamiento constante, mientras que la Prevención Operativa funciona como un implementador de las estrategias de seguridad.

En cambio, las Medidas con Terceros y Gobernanza de Seguridad garantizan que los estándares de seguridad no solo se implementen en el ámbito interno, sino que también se propaguen a toda la cadena de suministro y se adhieran a normativas internacionales.

En suma, estos subcomponentes aseguran un enfoque completo de ciberseguridad en el sector marítimo, garantizando que las amenazas sean identificadas y neutralizadas de forma eficaz.

La Tabla III presenta una comparativa entre marcos y modelos de ciberseguridad frente a criterios clave del entorno marítimo. Esta evaluación evidencia que los marcos tradicionales, como ISO/IEC, NIST CSF o COBIT, ofrecen una cobertura limitada cuando se aplican al dominio naval, aspecto que aborda de manera directa POSEIDON reflejando su diseño orientado a la realidad operativa naval y a la integración de elementos doctrinales y culturales clave para la resiliencia cibernética en el ámbito marítimo. Los criterios fueron definidos a partir de un análisis funcional del entorno marítimo-militar, teniendo en cuenta la necesidad de operar en entornos híbridos (físico-digitales), dar continuidad operacional y misional bajo ciberataques, incorporar una cultura de ciberseguridad en tripulantes embarcados, y facilitar la integración con doctrinas existentes, todo sin comprometer la operatividad.

Tabla III
Comparativa referentes de ciberseguridad frente a criterios del entorno marítimo

Marco / Modelo	Cobertura sectorial marítima	Entornos híbridos	Soporte por misión	Enfoque doctrinal adaptable	Cultura de ciberseguridad embarcada	Naturaleza del referente
ISO/IEC	Parcial	Parcial	No	No	No	Estándar de gestión
NIST CSF	Parcial	Sí	No	No	Parcial	Marco de referencia
MITRE ATT&CK	No	No	No	No	No	Matriz táctica
ENISA (puertos)	Sí (infraestructura)	Parcial	No	No	No	Guía técnica
OWASP	No	Parcial (web)	No	No	No	Proyecto abierto
COBIT	No	Parcial	No	Parcial	No	Marco de gobierno TI
BSI IT-Grundschutz	Parcial	Parcial	No	No	No	Catálogo de controles
POSEIDON	Sí (misión naval)	Sí	Sí	Sí	Sí	Marco sectorial adaptativo

V. ESTUDIO DE CASO: COMPONENTE DE PLANEACIÓN EN EL BUQUE DUAL

Este artículo, además de describir el marco POSEIDON, se centra en la fase de planificación, validando por primera vez sus subcomponentes y procesos en un caso de estudio real. Este tipo de embarcación representa un desafío adicional en ciberseguridad debido a la integración de sistemas avanzados de monitoreo, navegación digital, investigación marítima y registro de datos digitales de equipos de exploración, lo que lo convierte en un objetivo atractivo para ataques cibernéticos ver Fig.3.

Este caso de estudio se desarrolló en las instalaciones del astillero, en donde se utilizó todo lo descrito en las JNIC2024 para analizar los riesgos cibernéticos y aplicar medidas correctivas y preventivas.

El buque de uso dual que opera en rutas internacionales, realizando diferentes tipos de levantamientos, análisis de muestras y diferentes actividades de investigación. Depende de sistemas de navegación GPS, comunicaciones satelitales y sistemas de gestión de manejo de datos, análisis de muestras y conservación de elementos, entre otros, y todos interconectados.

En 2023, el buque según los reportes del centro de protección fue víctima de un ciberataque que comprometió uno

de los dos sistemas GPS afectando sus sistemas de navegación y colocando en riesgo la seguridad de la tripulación y los equipos. Para el objetivo de la divulgación académica, se realiza un análisis de todos los registros del evento para luego ser comparados al someter el buque al mismo evento con los parámetros del marco de trabajo POSEIDON. En este escenario, el componente de planificación en el marco POSEIDON desempeña un rol vital, al facilitar la creación y organización de estrategias de seguridad, detectar vulnerabilidades y definir respuestas eficaces ante incidentes de ciberseguridad.

A. Identificación del Problema Basada en las Amenazas del Buque Dual.

Mediante el estudio del buque dual presentado en la JNIC2024, se detectaron varias amenazas que están resumidas en la Tabla IV y que podrían comprometer su seguridad en las operaciones [19]. Estas amenazas se clasificaron y trataron a través de estrategias de planificación en POSEIDON, garantizando que cada una presentara una respuesta organizada. estrategias de planificación en POSEIDON, garantizando que cada una presentara una respuesta organizada.

Tabla IV
IDENTIFICACIÓN DE AMENAZAS DEL BUQUE DUAL

Familia de Amenazas	Tipo de Amenazas	Impacto y estrategias de mitigación
Desastres naturales, tecnológicos	Fallas tecnológicas y ambientales.	Sistemas de respaldo y redundancia para garantizar la continuidad operativa.
Interrupciones	Fallos en servidores, ataques a propulsores.	Plan de contingencia con servidores alternativos y protocolos de respuesta rápida.
Daño no intencional	Errores humanos, fuga de información.	Capacitaciones en ciberseguridad y control de accesos con autenticación multifactor.
Agresiones físicas	Piratería, hacktivismo.	Integración de seguridad física y digital para detectar ataques combinados.
Fallos en sistemas críticos	Fallos en GPS y mesa de cartas.	Sistemas de respaldo y redundancia en equipos de navegación.

B. Implementación Subcomponentes en el Caso de Estudio

La implementación de los subcomponentes del componente de planificación en el caso de estudio del buque dual permitió hacer una primera validación de su aplicabilidad en un entorno operativo real. Cada subcomponente fue abordado con un enfoque práctico y adaptado a las condiciones del buque que era objeto del caso de estudio. A continuación, se describe cómo se aplicaron estos subcomponentes en el buque, destacando los resultados observados y las mejoras derivadas del proceso de planificación estructurada:

- Prevención: Durante toda la etapa de diseño del buque, se implementó un plan de ciberseguridad que realizaba una vigilancia constante de todas las vulnerabilidades del buque y cada uno de sus sistemas. Se detectaron mecanismos de acceso no autorizado a los sistemas de automatización del motor, y algunos equipos de los sistemas de investigación de los científicos de abordo lo que llevó a reforzar los controles de autenticación y la implementación de jornadas de concienciación a bordo de la Unidad, entre otras acciones.
- Perfil del Personal: Se evaluó previamente el nivel de conocimiento de la tripulación en ciberseguridad mediante encuestas y pruebas de análisis de conocimiento y comportamiento. Esto se realizó con la asesoría de profesionales en psicología especializados en estrategias para fortalecer la cultura de ciberseguridad en el factor humano. La mecánica realizada permitió, identificar brechas en la gestión de accesos y respuesta a incidentes. Como resultado, se implementaron capacitaciones periódicas con diferentes estrategias lúdicas y del aprendizaje. De igual manera se implementó un protocolo de seguridad para el uso de dispositivos personales a bordo.
- Lecciones Aprendidas: Durante las primeras operaciones del buque, se registraron fallas en la comunicación entre los sistemas de propulsión y monitoreo ambiental debido a configuraciones inseguras en la red. Esta experiencia permitió optimizar la segmentación de redes y establecer reglas más estrictas para la comunicación entre sistemas críticos.
- Prevención Operacional: Se integró un sistema de detección de intrusiones en la red del buque, lo que permitió identificar patrones de tráfico anómalos antes de que se materializara una amenaza. Como parte del proceso de mejora, se implementaron actualizaciones periódicas de software y pruebas de penetración simuladas para evaluar la resiliencia del buque ante ciberataques.
- Medidas con Terceros: Se realizó una auditoría de ciberseguridad a los proveedores de software y hardware del buque, detectando que algunos sistemas no cumplían con las certificaciones requeridas. Como consecuencia, se establecieron nuevos requisitos para proveedores y la obligación de pruebas de seguridad antes de cualquier integración con los sistemas del buque.
- Gobernanza de Seguridad: Se formalizó un marco de gobernanza de ciberseguridad que incluyó procedimientos de respuesta a incidentes, monitoreo regulatorio y coordinación con agencias internacionales. Esto permitió al buque cumplir con normativas de seguridad marítima como las directrices de la OMI sobre ciberseguridad en buques.

C. Evaluación de Resultados Cuantificables en la Resiliencia Cibernética

Frente al ataque al que fue sometida la Unidad Marítima, en el análisis original se menciona que la resiliencia cibernética del buque mejoró en un 60% tras la implementación del marco POSEIDON. Los investigadores estimaron este dato mediante métricas concretas para evaluar el impacto de la ciberseguridad en el buque dual analizado:

- Tiempo Promedio de Recuperación (MTTR, Mean Time to Recovery): Se realizó una comparación del tiempo de recuperación de sistemas críticos antes y después de la implementación de POSEIDON. Antes de la aplicación del marco, según los registros encontrados de eventos anteriores, el tiempo promedio de restauración de sistemas básicos afectados por ciberataques era de 48 horas. debido a que se requería el acceso de personal experto con el software de restauración y no se contaba con backups oportunos. Tras la implementación de POSEIDON, el tiempo bajo a 24 horas, lo que indica que obtuvo una mejora en promedio del 50% en la capacidad de recuperación del sistema de a bordo del buque.
- El Tiempo Promedio de Detección de Incidentes (MTTD, Mean Time to Detect): Antes del uso de POSEIDON, la detección de ataques como intrusiones en los sistemas de navegación o intentos de manipulación del AIS (Automatic

Identification System) promediaba 6 horas ya que estaba directamente relacionado con la detección por intervención humana. Con la implementación de procesos de monitoreo continuo y detección temprana de incidentes con herramientas validadas vía software, este tiempo disminuyó a 2 horas, logrando una mejora en la capacidad de respuesta de un 67%.

- Reducción de Incidentes de Phishing y Ransomware: Por medio de actividades dirigidas a la tripulación como actividades de sometimiento de estrés (simulacros), capacitaciones y la adopción de controles de seguridad avanzados, se redujeron los incidentes positivos en un 40%, según reportes basados en los registros de dispositivos de protección como el firewall y controles vía software del servidor. En términos de ataques de ransomware, la combinación de segmentación de redes y autenticación de doble factor en todos los servicios permitió disminuir la efectividad de estos ataques en un 55%. Es importante resaltar la disposición que se pudo observar en el personal para aportar a los procesos que se ejecutaban para mitigar riesgos latentes de ataque luego de participar en las actividades por estrés donde además de simular pérdida de capacidades tecnológicas se evaluó las reacciones del factor humano [20-24]. Las métricas clave permiten la valoración de la eficacia de los simulacros al medir los tiempos de respuesta, exactitud en la realización, coordinación del equipo y grados de estrés. Estos indicadores facilitan la detección de fallos, la mejora de protocolos y la capacitación del personal frente a incidentes reales, asegurando una reacción más eficaz y segura. Las métricas usadas se registran en la Tabla V.
- Impacto en la Disponibilidad de Sistemas Críticos: Se evidenció un aumento del 30% en la disponibilidad operativa de los sistemas de navegación y comunicaciones debido a los tiempos de respuesta, verificación de arquitectura, validación de tipos de acceso a los equipos y a una mayor robustez en la arquitectura de seguridad y respuesta rápida a incidentes.
- Eficiencia en la Coordinación con Terceros: Con la mejora en los protocolos de seguridad con proveedores y operadores de puertos, se redujo en un 20% el tiempo de validación de integridad de sistemas externos conectados al buque, con la implementación de procesos previos de validación de personas, verificación de equipos, limpieza de hardware que era usado en diferentes actividades, entre otras medidas de control.

D. Lecciones Aprendidas y Mejoras Implementadas

A partir del análisis del buque dual, se extrajeron varias lecciones clave que además de enriquecer el proceso de investigación en general, han permitido optimizar el componente de Planificación dentro del marco de trabajo POSEIDON:

- Anticiparse es clave en todos los procesos de ciberseguridad:
  Desarrollar una planificación proactiva permite detectar amenazas y vulnerabilidades antes de que conviertan en incidentes críticos.
  El análisis de todos los activos críticos es fundamental para establecer estrategias de protección adecuadas y dimensionar la curva de disminución de capacidades.
- La concienciación y capacitación de los tripulantes es básica:
  Se constató la necesidad de concienciación y se evidenció que la falta de formación en ciberseguridad está directamente relacionada con el aumento del riesgo de errores humanos.
  Un plan de choque acompañado de la realización de entrenamientos periódicos mejoró la capacidad de respuesta ante amenazas.
- Implementar medidas de seguridad física y digital es básico:
  La combinación de seguridad cibernética con cualquier medida de restricción y/o protección física contribuye a la reducción del impacto de amenazas híbridas (como ataques combinados de piratería y ciberataques).
- La vigilancia y monitoreo constante de todas las estrategias es necesaria por el continuo avance de las tecnologías:
  La evolución constante de las amenazas requiere una actualización dinámica de los protocolos de seguridad.
  La revisión periódica de la planificación permitió mejorar los procedimientos de respuesta a incidentes.
- La colaboración con terceros debe ser más rigurosa:
  La auditoría de proveedores reveló vulnerabilidades en algunos sistemas externos integrados al buque.
  Se establecieron requisitos más estrictos para garantizar la seguridad en la cadena de suministro.

TABLA V
MÉTRICAS USADAS PARA EVALUAR EL FACTOR HUMANO

Categoría	Métrica	Descripción
Tiempo de Respuesta	Tiempo de reconocimiento	Tiempo desde la alerta hasta la primera reacción.
	Tiempo de decisión	Cuánto tarda el equipo en tomar una acción.
	Tiempo de ejecución	Tiempo total para completar la respuesta.
Precisión y Efectividad	Errores cometidos	Número de errores en la ejecución del protocolo.
	Cumplimiento del protocolo	% de acciones realizadas según el plan establecido.
Coordinación y Comunicación	Número de fallas en la comunicación	Interrupciones o malentendidos en la coordinación.
	Fluidez del trabajo en equipo	Evaluación subjetiva de la colaboración.
Estrés y Reacciones Fisiológicas	Nivel de estrés	Medido con sensores biométricos o cuestionarios.
	Seguridad percibida	Autoevaluación sobre confianza en la respuesta.

VI. CONCLUSIONES

El marco POSEIDON ha demostrado ser una herramienta efectiva para fortalecer la ciberseguridad en el sector marítimo, proporcionando una estructura robusta para la identificación de riesgos, planificación estratégica y optimización de la respuesta ante incidentes. Por medio del caso de estudio presentado, se pudo mostrar como la integración de los procesos clave como monitoreo, análisis de amenazas y coordinación con terceros, entre otros que contribuyen al incremento de la resistencia cibernética de la infraestructura la marítima.

Este estudio expande el análisis presentado en [1] refinando la validación del marco POSEIDON con nuevas métricas y una evaluación más profunda de su fase de planificación. Además, incorpora un enfoque más detallado en la integración del marco con las necesidades operativas del sector marítimo. Los resultados brindan una base sólida para que sea aplicado en otros tipos de infraestructuras marítimas y se sigue confirmando la efectividad del modelo.

Asimismo, se ha planteado la evaluación integral de todas las fases de POSEIDON en futuras investigaciones, garantizando su aplicación en diversos escenarios operativos y explorando oportunidades de optimización en distintos entornos.

El estudio reafirma la importancia de una ciberseguridad

proactiva, donde la mejora continua, la supervisión en tiempo real y una sólida cultura de seguridad sean elementos fundamentales. A medida que evolucionan las amenazas cibernéticas, POSEIDON debe continuar su proceso de adaptación, incorporando nuevas metodologías y alineándose con estándares internacionales emergentes. Su flexibilidad y capacidad de actualización lo posicionan como un modelo clave en la protección de infraestructuras marítimas.

Este enfoque estructurado no solo facilita la identificación de áreas de mejora, sino que también proporciona una base sólida para la implementación y evolución continua del marco POSEIDON en un entorno marítimo cada vez más complejo y expuesto a amenazas cibernéticas.

Líneas de Trabajo Futuro: El estudio ha demostrado el impacto positivo de la fase de planificación de POSEIDON en la ciberseguridad marítima, sentando las bases para su aplicación en infraestructuras más amplias. El siguiente paso será la validación integral de todas las fases de POSEIDON en diversos entornos operativos, garantizando su aplicabilidad en buques militares, comerciales y de investigación. Se emplearán pruebas de campo y simulaciones avanzadas para evaluar la interacción entre las fases del marco y su impacto en la resiliencia cibernética del sector marítimo.

Otro eje clave será la gestión del factor humano, incluyendo entrenamientos avanzados y el análisis del comportamiento en simulacros de ciberataques. También se reforzarán los mecanismos de supervisión, auditoría y coordinación con proveedores externos, garantizando estándares más estrictos en la cadena de suministro. POSEIDON continuará evolucionando mediante evaluaciones conjuntas de todas sus fases, asegurando su capacidad de adaptación a un entorno marítimo dinámico y alineándose con normativas internacionales para fortalecer la seguridad cibernética a nivel global.

AGRADECIMIENTOS

Este trabajo se ha desarrollado gracias a la financiación proporcionada por los siguientes proyectos: Proyecto Di4SPDS (CHIST-ERA grant-PCI2023145980-2) financiado por MCIN/AEI y cofinanciado por la Unión Europea, KOSMOS-UCLM (PID2024-155363OB-C44) financiado por MCIN/AEI/10.13039/501100011033; ALBA-UCLM (TED2021-130355B-C31) financiado por CIN/AEI/10.13039/501100011033/ Unión Europea NextGenerationEU/PRTR, y MESIAS (2022-GRIN-34202) financiado por FEDER, así como por la Corporación de Ciencia y Tecnología para el Desarrollo de la Industria Naval, Marítima y Fluvial, COTECMAR. finalmente, agradecer al personal de la Armada Nacional de Colombia que participó en este trabajo, por su disposición y compromiso.

REFERENCIAS

[1] Y. Zhang, H. Zhao, R. Jiang, and Y. Shang, "Manage morning commute problem of household travels under single-step toll: A comparison study," *Transp Res Part C Emerg Technol*, vol. 174, May 2025, doi: 10.1016/j.trc.2025.105075.

[2] D. Faz, A. O.-C. e Ingeniería, and undefined 2014, "Revisión del estado del arte de la evaluación de riesgos en instalaciones portuarias.," *redalyc.org*, vol. 35, no. 2, pp. 85–93, 2014, Accessed: Mar. 19, 2025. [Online]. Available: https://www.redalyc.org/pdf/5075/507550625004.pdf

[3] J. Anabalón, E. D.-E. de S. y Defensa, and undefined 2014, "Una revisión de ciberdefensa de infraestructura crítica," *anepe.cl*, Accessed: Mar. 19, 2025. [Online]. Available: https://anepe.cl/wp-content/uploads/2024/05/ESD_3.pdf#page=132

[4] R. Lewis-Fernández, Hinton Devon E, Laria Amaro J., Patterson Elissa H, Hofmann Stefan G, Craske Michelle G, Stein Dan J, Asnaani Anu and Betty Liao Ma, "Culture and the anxiety disorders: recommendations for DSM-V," *psychiatryonline.org*, Accessed: Mar. 19, 2025. [Online]. Available: https://psychiatryonline.org/doi/abs/10.1176/foc.9.3.foc351

[5] J. C. Crawford, "Ciberataque Al Transporte Marítimo. ¿Una Amenaza Real O Ciencia Ficción?," *Revista de Marina Nº*, vol. 970, pp. 15–23, Jun. 2019, Accessed: Feb. 27, 2022. [Online]. Available: https://revistamarina.cl/revistas/2019/3/jcrawfordc.pdf

[6] W. C. Leite Junior, C. C. de Moraes, C. E. P. de Albuquerque, R. C. S. Machado, and A. O. de Sá, "A Triggering Mechanism for Cyber-Attacks in Naval Sensors and Systems," *Sensors (Basel)*, vol. 21, no. 9, May 2021, doi: 10.3390/S21093195.

[7] F. Martínez, L. E. Sánchez, A. Santos-Olmo, D. G. Rosado, and E. Fernández-Medina, "Maritime cybersecurity: protecting digital seas," *Int J Inf Secur*, vol. 23, no. 2, pp. 1429–1457, Apr. 2024, doi: 10.1007/S10207-023-00800-0/TABLES/7.

[8] A. Androjna, M. Perkovič, I. Pavic, and J. Mišković, "AIS data vulnerability indicated by a spoofing case-study," *Applied Sciences (Switzerland)*, vol. 11, no. 11, Jun. 2021, doi: 10.3390/APP11115015.

[9] A. Amro and V. Gkioulos, "From Click to Sink: Utilizing AIS for Command and Control in Maritime Cyber Attacks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13556 LNCS, pp. 535–553, 2022, doi: 10.1007/978-3-031-17143-7_26/COVER.

[10] A. Altaweel, H. Mukkath, and I. Kamel, "GPS Spoofing Attacks in FANETs: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 55233–55280, 2023, doi: 10.1109/ACCESS.2023.3281731.

[11] W. C. Leite Junior, C. C. de Moraes, C. E. P. de Albuquerque, R. C. S. Machado, and A. O. de Sá, "A Triggering Mechanism for Cyber-Attacks in Naval Sensors and Systems," Sensors (Basel), vol. 21, no. 9, May 2021, doi: 10.3390/S21093195.

[12] B. Svilicic, J. Kamahara, J. Celic, and J. Bolmsten, "Assessing ship cyber risks: a framework and case study of ECDIS security," *WMU Journal of Maritime Affairs*, vol. 18, no. 3, pp. 509–520, Sep. 2019, doi: 10.1007/S13437-019-00183-X/FIGURES/6.

[13] Santos-Olmo, A., Sánchez, L. E., Rosado, D. G., Serrano, M. A., Blanco, C., Mouratidis, H., and Fernández-Medina, E., "Towards an integrated risk analysis security framework according to a systematic analysis of existing proposals," *Front Comput Sci*, vol. 18, no. 3, pp. 1–18, Jun. 2024, doi: 10.1007/S11704-023-1582-6/METRICS.

[14] ISO, "ISO - ISO 16425:2013." [Online]. Available: https://www.iso.org/standard/56739.html

[15] ISO, "ISO/IEC 27001:2022 - Information security management systems," 3. Accessed: Jul. 10, 2024. [Online]. Available: https://www.iso.org/standard/27001

[16] ENISA, "Guidelines - Cyber Risk Management for Ports — ENISA." Accessed: Jul. 26, 2023. [Online]. Available: https://www.enisa.europa.eu/publications/guidelines-cyber-risk-management-for-ports

[17] ENISA, "Agencia de la Unión Europea para la Ciberseguridad - ENISA." Accessed: Mar. 27, 2023. [Online]. Available: https://www.enisa.europa.eu/

[18] NIST, "The NIST Cybersecurity Framework (CSF) 2.0," Feb. 2024, doi: 10.6028/NIST.CSWP.29.

[19] S. Ordás Jiménez, I. Bazán García, and R. Santalices Fernández, "La Protección de los sectores marítimo y portuario: amenazas y respuestas consideradas por la Organización Marítima International," *La Protección de los sectores marítimo y portuario: amenazas y respuestas consideradas por la Organización Marítima International*, Jan. 2012, doi: 10.5821/EBOOK-9788476539613.

[20] Colegio de Psicólogos, "Test psicológico stai: cómo se califica y qué es." Accessed: Nov. 20, 2024. [Online]. Available: https://colegiodepsicologossj.com.ar/test-psicologico-stai/

[21] Ana Teresa Carbonel and Antonio Torres, "Evaluación de percepción de riesgo ocupacional." Accessed: Nov. 20, 2024. [Online]. Available: http://scielo.sld.cu/scielo.php?pid=S1815-59442010000300003&script=sci_arttext&tlng=en

[22] F. Sánchez Vera, J. E. Martínez Guirao, and A. Téllez Infantes, "La seguridad en el ciberespacio desde una perspectiva sociocultural," *methaodos revista de ciencias sociales*, vol. 10, no. 2, pp. 243–258, oct. 2022, doi: 10.17502/MRCS.V10I2.577.

[23] S. G. Lilian and C. S. Andrea, "Confiabilidad y validez de la escala de impulsividad de Barratt (BIS-11) en adolescentes," *Rev Chil Neuropsiquiatr*, vol. 51, no. 4, pp. 245–254, 2013, doi: 10.4067/S0717-92272013000400003.

[24] R. Montañez, E. Golob, and S. Xu, "Human Cognition Through the Lens of Social Engineering Cyberattacks," *Front Psychol*, vol. 11, p. 528099, Sep. 2020, doi: 10.3389/FPSYG.2020.01755/BIBTEX.

Sesión Investigación V: Criptografía clásica y post-cuántica

Exploration of Deep Learning Convolutional Models Applied to Side-channel Analysis

David Perez Perez
Ikerlan, S. Coop
Arrasate-Mondragon, España
dperez@ikerlan.es

Servio Paguada
Ikerlan, S. Coop
Arrasate-Mondragon, España
slpaguada@ikerlan.es

***Abstract*—**
Cryptographic algorithms such as AES have been demonstrated to be robust enough to ensure secure communications. Distributed hardware devices, particularly IoT systems, are increasingly vulnerable to side-channel attacks. These attacks exploit information leaked by hardware during cryptographic operations to extract protected keys. Deep learning techniques have significantly enhanced the effectiveness of such attacks, yet their full potential remains uncertain, complicating security assessments. In this paper, we explore deformable convolutional models as a potentially more effective approach for conducting side-channel attacks. Additionally, we propose a new evaluation metric to assess the performance of these models, aiming to support the development of improved security countermeasures.

***Index Terms*—Side-channel attack, CNN, deformable-CNN, side-channel evaluation metrics**

Type of contribution: *Original research*

I. Introduction

The symmetric encryption algorithm AES has been the standard since 2002 [1]. In recent years, with the proliferation of IoT devices, there has been an increase in the risk associated with side-channel attacks [2], [3]. These types of attacks are based on the analysis of side-channel information, which encompasses all forms of information that a hardware device can emit while executing a cryptographic algorithm, such as consumed power, electromagnetic emissions, or sound. This information is indirectly related to the sensitive data handled by the device, meaning that an adequate analysis of the side-channel information has the potential to leak confidential information, thereby creating a breach in communication security.

Countermeasures applicable to cryptographic algorithms were developed to ensure the confidentiality of the processes executed on devices. These techniques allowed the decorrelation of the side-channel information from the sensitive information. Nevertheless, since they remained indirectly related, it was observed that it was still possible to recover sensitive information using convolutional neural networks (CNN) [4]. These were used to tackle the problem as a classification task, where each signal is associated with a key.

The incorporation of CNN has brought some challenges when evaluating the security of hardware devices against side-channel attacks. On one hand, the maximum information recovery capability of neural network models is unknown. This has motivated the search for more capable models that help defining the limits of these attacks [5], [6], [7]. On the other hand, the artificial intelligence models used in the attacks cannot be evaluated using common functions such as accuracy and cross-entropy. Those metrics do not adequately assess the effectiveness of the models when having a group of signals that are leaking the same private data. Contrary, the use of the guessing entropy metric was standardized in the area of side-channel attacks [8]. Although it adequately evaluates the model's effectiveness, its high computational cost and representation make it harder to use during the training phase. Therefore, researchers continue to search for informative metrics on the effectiveness of the attacks in studies such as [9], [10], [11].

The study of new deep learning architectures is an active topic in the field, as several of those have not yet been used to conduct side-channel analysis evaluation. In this paper, we present the deformable convolution, a type of CNN convolution capable of adapting its filters to the input signal, allowing deeper analysis [12].

The contributions of this paper are:

- We propose deformable convolutional networks as a potential improvement over the current state-of-the-art in the context of side-channel analysis.
- We introduce a metric called convergence for evaluating neural network models within the area of side-channel analysis.

The sections of this paper are the following: The background section explains side-channel attacks, convolutional neural networks, and the guessing entropy metric. Deformable convolutional models section introduces the architecture we propose. Convergence metric explains and analyses the properties of the metric we propose. Experimental results section explain the experiments and discusses our results.

II. Background

A. Side-channel Attacks

Side-channel attacks are a class of attacks that create security breaches by exploiting physical emissions, commonly known as traces, leaked by hardware devices during the execution of cryptographic algorithms. Introduced by Kocher et al. [2], [3], side-channel attacks take advantage of involuntary emissions such as sound, power consumption, and electromagnetic radiation.

Assuming the attacker has full control over an identical device to the target, he can perform a profiling side channel attack, one of the most dangerous type of side-channel attack. These are typically executed in two phases. (1) Profiling phase, during which the attacker builds a model capable of recovering sensitive information from the traces. (2) Attack

phase, the attacker captures traces from the target device and analyzes them with the previously acquired model to extract the protected information. Traditionally, the template attacks were ones of the most proficient attacks that could be used [13].

For example, these attacks can recover the private keys employed in communications secured by AES. To accomplish this, the output from the first AES's SubBytes phase is taken as the sensitive intermediate value to be extracted from the traces. This value is known as the intermediate value and is related to the private key through Eq. (1), where p represents the plaintext, which is combined via an XOR operation with k, the key. The S-box is a nonlinear function integral to AES that was chosen as the target point because of its strong correlation with k and the significant information leakage it generates in the traces, rendering it exploitable.

$$y = \text{S-box}(p \oplus k) \tag{1}$$

Countermeasures are used to mitigate the effectiveness of side-channel attacks. These countermeasures reduce the direct correlation between the traces and the intermediate values. Examples of countermeasures include executing certain AES operations in random order and applying masking techniques to the algorithm's variables using mask values [4]. In the case of the intermediate values, masking operates as illustrated in Eq. (2).

$$y = \text{S-box}(p \oplus k) \oplus m \tag{2}$$

B. Convolutional Neural Networks

A convolutional neural network (CNN) is a type of deep learning model widely used in signal analysis due to its ability to extract local features from the input data. This model is characterized by the use of convolutional layers, where a convolution operation is performed between the input and a kernel composed of trainable parameters. These parameters are optimized during training by minimizing a loss or objective function. Once the convolutional layer processes the input, the resulting output is passed through a nonlinear activation function, yielding a set of feature maps. These feature maps are subsequently processed by a pooling layer, which reduces the dimensionality of the signals through subsampling. This process produces a compact yet highly informative representation of the input. This compact output reduces the computational costs of later processing stages and grants the model invariance to small local transformations, such as signal shifts.

In Eq. (3), we present the convolution operation used in the convolutional layer. In this formula, x and W denote arrays representing the input signal to the layer and the filters, respectively. The dimensions of these arrays include the output feature index c_o, the input feature index c_i, the filter element index μ, and the input signal element index n. Notably, the latter indices also indicate the centers for the product between the filters and the input signal.

$$y^{c_o,n} = \sum_{c_i,\mu} W^{c_o,c_i,\mu}\ x^{c_i,n+\mu} \tag{3}$$

Convolutional neural network models have been applied to side-channel attacks, as these problems can be framed as classification tasks where the objective is to minimize the cross-entropy loss function. It has been demonstrated that CNNs offer advantages when analyzing traces, as they can extract more relevant information regarding intermediate values compared to traditional algorithms [4].

Currently, the research community is actively investigating the properties of the filters that yield the most efficient CNNs for side-channel attacks, given that filter design appears to have a significant impact on model performance. Some researchers, such as Zaid et al. [5], suggest that employing an initial layer with filters of size 1 could produce more effective models by potentially reducing the influence of noise on the output. Other researchers, like Wouters et al. [14], show that this statement is subject to doubt. Additionally, Paguada et al. [6] argue that using a large dilated filter might provide the model with an advantage. Dilated convolution enables outputs to be influenced by the most informative points of the intermediate values besides the mask used as the countermeasure, as this one also leaks information through the traces. This uncertainty regarding the optimal filter size in convolutional networks motivates our proposal of deformable convolutional models as a potential next state-of-the-art.

C. Guessing Entropy

Conventional metrics such as the loss function (commonly used for convolutional models) or accuracy are not optimal when assessing the performance of a model in recovering keys through side-channel attacks. In the field of side-channel analysis, we are primarily interested in how the model's ability to classify traces with intermediate values evolves as the number of analyzed traces increases. To capture this behavior, guessing entropy has become a standard [8].

Guessing entropy is a metric that estimates the average number of attempts an attacker would need, given a set of concurrently analyzed traces, to correctly guess the key, assuming the attempts are made in descending order of probability. This metric not only evaluates the model's capability to recover the key but also indicates the number of traces required for successful recovery.

A major drawback of guessing entropy is that its output is a graph depicting guessing entropy as a function of the number of traces—is challenging to utilize during the training phase. Visualizing and comparing guessing entropy curves from different stages of training can be difficult. To address this issue, we introduce a new metric called convergence.

III. Deformable Convolutional Networks

Deformable convolutional neural networks were introduced by Jufeng Dai et al. [12]. Unlike conventional convolutional models, these networks feature filters whose shapes are not fixed but dynamically adapt to the input data during training. In [12], Dai et al. demonstrate that deformable models can enhance image classification performance by increasing the focus on the most informative regions of the input through a more consistent alignment of filter elements with these key points. This property is especially appealing in the context of side-channel analysis, where emphasizing informative points

in intermediate values or masks may lead to improved performance. Various versions of deformable convolutional models exist. In our study, we will investigate dynamic models, static models, and their modulated variants.

The dynamic deformable model was developed by Jufeng. It follows the formula of Eq. (4), where the only apparent difference from a conventional convolution is the inclusion of $\Delta_d^{\mu,n}$, which defines a displacement for the point analyzed by the μ-th element of the filter relative to the center of signal n. The values $\Delta_d^{\mu,n}$ are computed as shown in Eq. (5), which performs a convolution on the input signal with as many output features c'_o as there are filter elements, thereby generating a distinct displacement for each element. Generating the displacements through convolution ensures that they are determined solely by the local properties of the input around the center n rather than by more distant values. This characteristic enables the model to adapt to the input through subtle perturbations of the filter shape.

$$y^{c_o,n'} = f\left(b^{c_o} + \sum_{c_i,\mu} W^{c_o,c_i,\mu} x^{c_i,n+\mu+\Delta_d^{\mu,n}}\right) \tag{4}$$

$$\Delta_d^{c'_o,n} = \sum_{c_i,\mu} W_d^{c'_o,c_i,\mu} x^{c_i,n+\mu} \tag{5}$$

Note that since $\Delta_d^{\mu,n}$ can be fractional, we compute $x^{c_i,n+\mu+\Delta_d^{\mu,n}}$ using linear interpolation from the adjacent input values.

The static deformable model was introduced by Liqiang Xiao et al. [15]. It employs the formula given in Eq. (6). Note that in this case, the displacements $\Delta_e^{c_o,\mu}$ do not depend on the input signal, as they do not vary with n. Instead, these are trainable variables with dimensions c_o and μ.

$$y^{c_o,n'} = f\left(b^{c_o} + \sum_{c_i,\mu} W^{c_o,c_i,\mu} x^{c_i,n+\mu+\Delta_e^{c_o,\mu}}\right) \tag{6}$$

Similarly to the dynamic models, $x^{c_i,n+\mu+\Delta_e^{\mu,n}}$ is computed via linear interpolation using the adjacent input values.

Since $\Delta_e^{\mu,n}$ does not depend on the input, Liqiang et. al. [15] recommended regularizing its training by using only the sign of the computed gradient for optimization. This is because the gradients of $\Delta_e^{\mu,n}$ would otherwise consider the entire signal, introducing noise. Moreover, this regularization, combined with the fact that these are trainable parameters, allows static deformable models to generate filters that are more globally adapted to the input.

Finally, we have modulated deformable models, which were introduced by Xizhou et al. [16]. These models employ the formula of Eq. (7). In this case, $\Delta m^{\mu,n}$ is obtained by applying a sigmoid function to a convolution over the input that produces μ features, as illustrated in the function of Eq. (8). This yields a set of values within the range $[0, 1]$. It is also noteworthy that this model is used to regularize a deformable model - either dynamic or static in our context - that generates Δ as specified in 7.

By multiplying the convolution elements by $\Delta m^{\mu,n}$, the contribution of each convolution element to the output is, on average, reduced. This smoothing of the displacement search space leads to the emergence of more diverse filter shapes.

$$y^{c_o,n'} = f\left(b^{c_o} + \sum_{c_i,\mu} W^{c_o,c_i,\mu} x^{c_i,n+\mu+\Delta} \Delta m^{\mu,n}\right) \tag{7}$$

$$\Delta_d^{c'_o,n} = f\left(\sum_{c_i,\mu} W_m^{c'_o,c_i,\mu} x^{c_i,n+\mu}\right) \tag{8}$$

IV. Convergence Metric

Convergence is a metric introduced in our work to objectively evaluate the performance of any model in the field of side-channel attacks by estimating the value to which the guessing entropy converges with an infinite number of traces.

Firstly, we require an analytical expression to approximate the guessing entropy function. An analysis of the mathematical properties of guessing entropy reveals that our expression should satisfy the following conditions:

- The expression must be differentiable, with positive range and domain.
- The function must be decreasing.

This is because guessing entropy is defined for a positive number of traces and, as illustrated in Section II-C, it is inherently a positive value. Regarding the second property, the function should be decreasing; specifically, if a model is capable of extracting relevant information from the traces, its guessing entropy will decrease monotonically until it converges to a value corresponding to the indistinguishable set of intermediate values. Note that this decreasing trend is based on the assumption that the model retains what it has learned from previous traces. Conversely, if the model fails to learn any useful information, it will output a uniform probability distribution over the classes of the intermediate value, which results in an expected guessing entropy of 128.5, as shown in Eq. 9, in the case of working with a byte with 256 classes.

$$ge(1) = \sum_{n=1}^{256} \frac{n}{256} = 128,5 \tag{9}$$

With these properties established, we propose the function of Eq. (10) as the approximation formula. In this expression, x represents the number of traces, s is the guessing entropy obtained from a single trace, n denotes the convergence value, and m and p are two positive parameters that generate a family of exponential functions with two degrees of freedom. We employ these functions to approximate the decreasing behavior of guessing entropy. As illustrated in Figure 1, the function successfully approximates an instance of guessing entropy.

$$ge(x) = n + (s - n)e^{m - mx^p} \tag{10}$$

Once we have the approximation, we can express the experimental guessing entropy values as $ge_{exp}(x) = ge(x) + r$ where r i s a random variable with zero mean and variance σ^2.

Using this approximation, we can estimate the integral of the guessing entropy up to a general value v.

$$\int_1^{v+1} dx ge(x) = nv + A(v) \tag{11}$$

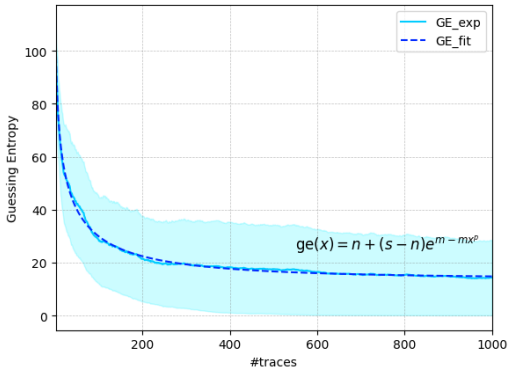


Fig. 1. Example of a guessing entropy obtained from an attack and fitted to the formula of Eq. (9).

$$\begin{aligned} A(\upsilon) &= (s-n)e^{m}\int_{1}^{\upsilon+1} dx e^{-mx^{p}} \\ &= \frac{(s-n)e^{m}}{pm^{\frac{1}{p}}}\left(\Gamma(\frac{1}{p},m)-\Gamma(\frac{1}{p},m(\upsilon+1)^{p})\right) \end{aligned} \tag{12}$$

We observe that the integral of the guessing entropy can be computed as the sum of two components: $n\upsilon$ and $A(\upsilon)$. Notice in Eq. (12) that $A(\upsilon)$ is an improper integral due to the inclusion of the incomplete Γ function. This function approaches zero for large values of υ, which means that $A(\upsilon)$ becomes increasingly independent of υ, eventually converging to a constant. We exploit this property by calculating the integral of the guessing entropy for a high value of υ over an interval $\Delta\upsilon$ to obtain an estimate of the limit at infinity of the guessing entropy, n, using our metric, called convergence $\hat{n}_2$.

$$\int_{\upsilon+1}^{\upsilon+\Delta\upsilon+1} dx ge(x) = n\Delta\upsilon + (A(\upsilon+\Delta\upsilon) - A(\upsilon)) \tag{13}$$

$$\hat{n}_2 = n + \frac{A(\upsilon+\Delta\upsilon) - A(\upsilon)}{\Delta\upsilon} \tag{14}$$

For a sufficiently large value of υ, the result of $\frac{A(\upsilon+\Delta\upsilon)-A(\upsilon)}{\Delta\upsilon}$ can be approximated as 0, which allows us to rewrite the convergence as:

$$\begin{aligned} \hat{n}_2 = n &= \frac{1}{\Delta\upsilon}\int_{\upsilon+1}^{\upsilon+\Delta\upsilon+1} dx ge(x) \\ \hat{n}_2 &\sim \frac{1}{\Delta\upsilon}\sum_{k=0}^{\Delta\upsilon} ge(\upsilon+k) \end{aligned} \tag{15}$$

By employing the convergence metric, we unlock the ability to analyze the value to which the attack's guessing entropy converges with an infinite number of traces. This enables us to study the model's maximum efficacy.

Once the metric is defined, we can examine its properties, such as its variance and bias relative to the estimated value n.

The variance of $\hat{n}_2$ can be computed using the following expression:

$$\begin{aligned} Var(\hat{n}_2) &= \frac{1}{\Delta\upsilon^2}\sum_{k=0}^{\Delta\upsilon} Var(ge(\upsilon+k)) \\ &= \frac{1}{\Delta\upsilon^2}\sum_{k=0}^{\Delta\upsilon} Var(r) = \frac{\sigma^2}{\Delta\upsilon} \end{aligned} \tag{16}$$

We can observe that by using the metric, integrating the guessing entropy values between υ and $\upsilon+\Delta\upsilon+1$ yields the divisor $\Delta\upsilon$. This represents a clear advantage compared to, for example, taking the guessing entropy value for a large number of traces as the convergence measure, since that value is random due to the random variable r, resulting in an estimate with variance σ^2 that is larger than that of our metric.

Regarding the bias of $\hat{n}_2$ relative to n, it arises from approximating $\frac{A(\upsilon+\Delta\upsilon)-A(\upsilon)}{\Delta\upsilon}$ by 0. To evaluate the accuracy of this approximation, we approximate $A(\upsilon)$ using a zero-order Taylor series expansion. By doing this, we can compute the Lagrange remainder of the series.

$$R(t) = |A(\upsilon+\Delta\upsilon) - A(\upsilon)| = \left|\frac{dA}{d\upsilon}(\upsilon+t\Delta\upsilon)\Delta\upsilon\right| \tag{17}$$

Note that t is a variable that can take any value between 0 and 1. To obtain $\frac{dA}{d\upsilon}$, we apply Leibniz's rule for differentiation under the integral sign, yielding:

$$\frac{dA}{d\upsilon} = (s-n)e^{m-m(\upsilon+1)^{p}} \tag{18}$$

Notice that since the derivative of $A(\upsilon)$ with respect to υ is an exponential function, it is always positive. This implies that $A(\upsilon)$ is monotonically increasing. Consequently, the difference $A(\upsilon+\Delta\upsilon) - A(\upsilon)$ is always positive. This observation indicates that our metric will act as an upper bound for n, as it is invariably greater than it.

Returning to the Lagrange remainder, we obtain:

$$\begin{aligned} \hat{n}_2 - n &= \sup_t \frac{R(t)}{\Delta\upsilon} \\ &= \sup_t (s-n)e^{m-m(\upsilon+t\Delta\upsilon)^{p}} \\ &= (s-n)e^{m-m\upsilon^{p}} \end{aligned} \tag{19}$$

From this, we deduce that the bias of the metric decreases exponentially with υ. Additionally, notice that it also takes values close to zero when $s \sim n$. This indicates that, for sufficiently large values of υ and $\Delta\upsilon$, the metric's bias will be nearly zero during the early stages of model training, when the guessing entropy functions oscillate around 128.5. Later, if the model manages to extract useful information from the traces, the metric's bias will increase until reaching a maximum, after which it will decrease back toward zero as m and p see their values being increased.

V. Experimental results

For the experiments, we utilized the ASCAD datasets created by Prouff et al. [4] using an ATMEGA8515 microcontroller executing AES_{128}. The ASCAD database is provided in two versions: ASCADf and ASCADr, which correspond to using a fixed key or random keys during the profiling phase. In ASCAD, the measurement windows of the raw traces are selected based on their highest correlation with the third byte

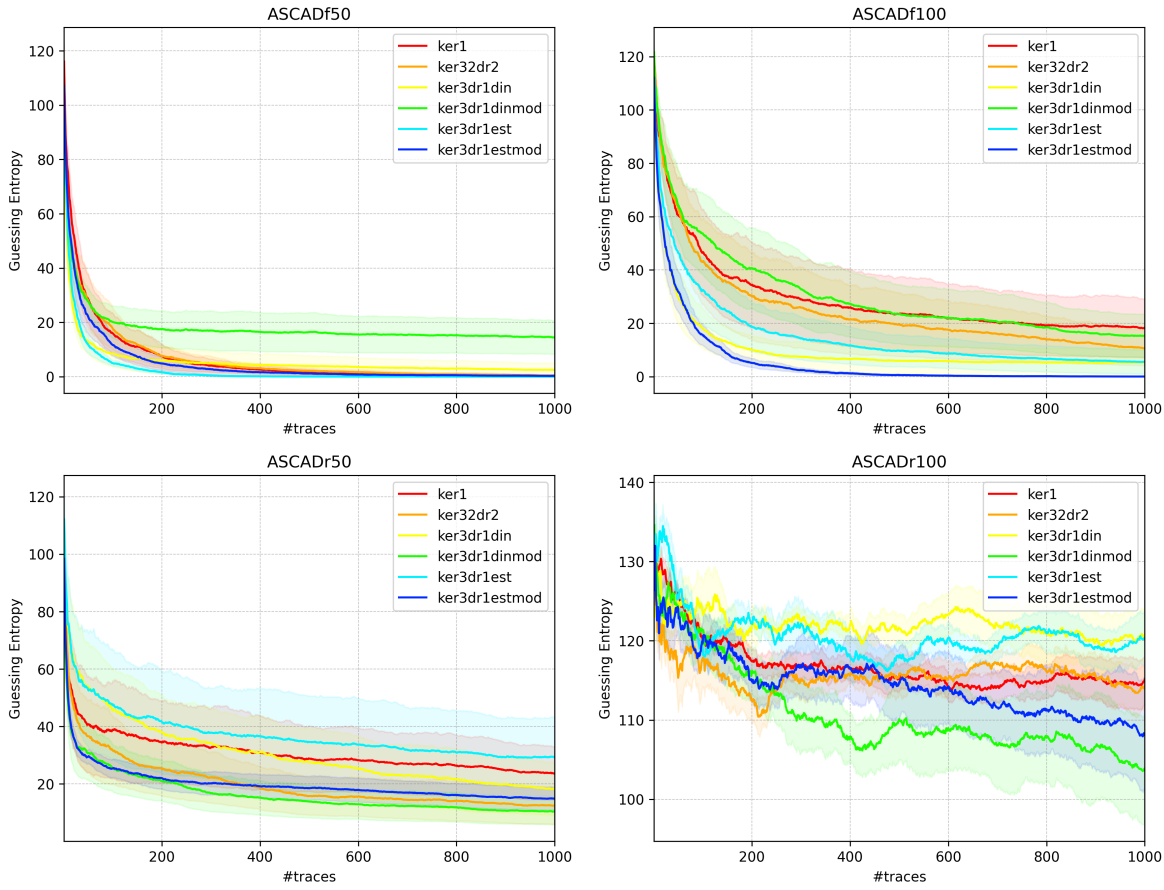


Fig. 2. Mean and standard deviation of guessing entropy functions for various models across different datasets.

of the masked intermediate values. Additionally, there are versions of the datasets where the traces are desynchronized by up to 0, 50, or 100 sample points - this is indicated by the numerical suffix appended to the dataset name. For example, ASCADr50 refers to the ASCADr dataset with offsets of up to 50.

The models used in our experiments are those from [5] for ASCADf and [6] for ASCADr, except that we replaced the first convolutional layer. This modification was made to study the impact of using a deformable first convolutional layer on performance compared to state-of-the-art models. Depending on the configuration of the first layer, we have:

- **ker1**: Convolutional layer with filters of size 1, proposed in [5].
- **ker32dr2**: Convolutional layer with filters of size 32 and dilation rate of 2, proposed in [6].
- **ker3dr1est**: Static deformable convolutional layer with filters of size 3.
- **ker3dr1estmod**: Modulated static deformable convolutional layer with filters of size 3.
- **ker3dr1din**: Dynamic deformable convolutional layer with filters of size 3.
- **ker3dr1dinmod**: Modulated dynamic deformable convolutional layer with filters of size 3.

Finally, we employed the entire ASCADf dataset and 30% of the ASCADr dataset for training. We set a maximum learning rate of $5E-3$, modulated using the OneCycle policy scheduler where we increase the learning rate during 40% of the training and use a div_factor of 10. The batch size used was 256, and the training was conducted over 50 epochs for ASCADf, 75 epochs for ASCADr0 and ASCADr50, and 100 epochs for ASCADr100. We select the third byte of the intermediate value as the classification label of the trace.

A. Evaluation of Deformable Convolution Networks

To evaluate the performance of the deformable convolutional models relative to the current state-of-the-art, we trained the models on the ASCADf50, ASCADf100, ASCADr50, and ASCADr100 datasets, obtaining the results shown in Figure 2.

As observed, the deformable models yield guessing entropy curves that are comparable or, in some cases, better than the state-of-the-art models, with some curves converging to lower values. This improvement is especially notable for the ASCADf100 dataset, where the three models with the lowest convergence values are ker3dr2estmod, ker3dr2est, and ker3dr2din.

In the experimental results, the regularizing effect of modulation is clearly observed. When training the models with simpler data from ASCADf, modulation appears to impede the performance of the deformable models, yielding guessing entropy curves that are either similar to or converge more slowly than those of unmodulated models. Conversely, when the models are trained on more complex data from

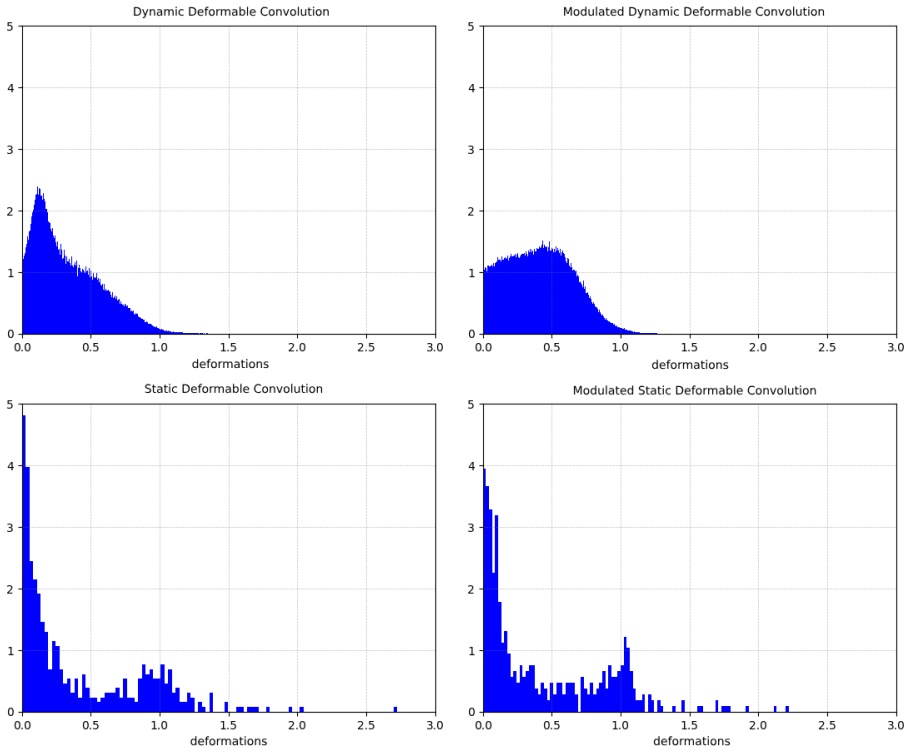


Fig. 3. Probability density functions of the absolute deformation values in the deformable convolutional layer trained with ASCADr0.

ASCADr, modulation seems to enhance the performance of the modulated models. This improvement is particularly evident in analyzing the guessing entropy of ker3dr2dinmod, which shows a substantial performance boost on ASCADr50 compared to earlier configurations.

When comparing the performance of the unmodulated deformable models, it can be observed that the dynamic model exhibits better guessing entropy curves, showing convergence values that are equal to or lower than those of the static model, except in the case of ASCADf50 where the static model achieves a lower convergence value of the guessing entropy. Moreover, when examining their modulated versions, ker3dr2estmod emerges as the best-performing model among those studied. This superiority is attributable to its consistent ranking among the models with the lowest convergence values across all datasets. Additionally, it is noteworthy that this model presents the smallest deviation values, indicating a low variability in its expected performance for each case.

Despite the apparent improvement in the performance of models like ker3dr2estmod, still fail to execute successful attacks in the ASCADr100 case. In this scenario, the complexity of the data combined with the desynchronization of 100 has led most models to exhibit guessing entropy curves oscillating around 120. Although the modulated deformable models display lower convergence values - typically between 110 and 100- these values are insufficient to classify the attacks as successful. It is also important to note that the deformable convolutional models require more computations, resulting in increments of up to 6 times in the time needed to complete an epoch.

Deformable convolutional models have been proposed as a potential enhancement, primarily due to their purported ability to adapt the size and shape of filters to the data. As previously mentioned, we hypothesize that this flexibility may be advantageous in executing more effective attacks, especially since filter shape appears to be a highly relevant variable according to prior research [5], [6]. This observation prompted us to explore the capacity to modify the size and shape of deformable models. We trained the deformable models on ASCADr0 to investigate this while recording the filter deformation values. This approach enabled us to generate probability density functions for the absolute values of the deformations, as shown in Figure 3.

Analyzing the results shown in Figure 3, we observe that the probability of encountering a deformation with an absolute value greater than one is significantly lower than that of obtaining a deformation below one. Let us consider a deformation with an absolute value lower than one as a perturbation of the filter's shape. It is evident that, in most cases, deformable models just introduce perturbations affecting both the shape and size of the filters.

When comparing the probability densities of the dynamic and static deformable convolutional models, it becomes evident that the dynamic variant tends to produce more non-zero deformations than the static one. Conversely, although the static model exhibits a peak probability at zero, it also shows a higher likelihood of generating deformations with absolute values greater than one compared to its dynamic counterpart.

Finally, when calculating the expected value of the absolute deformations, we observe that using a modulation leads to larger overall deformations, allowing the expected values to be increased from 0.3364 to 0.4133 in the dynamic deformable

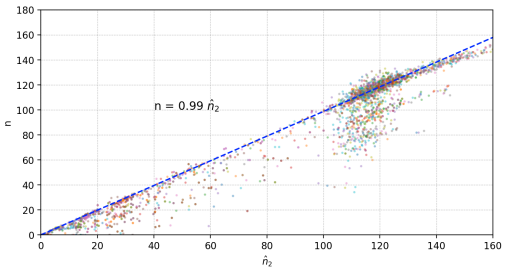


Fig. 4. Linear regression with weighted least squares of n y $\hat{n}_2$.

model and from 0.4129 to 0.4351 in the static one. This makes sense with the theory as the modulated versions should provoke a reduction in the relevance appointed to the output by each filter element, allowing it to get more deformed like in [16].

B. Evaluation of Metric Convergence

To assess the metric's effectiveness in indicating the convergence value of a model's guessing entropy, we computed this value alongside the convergence limit of the guessing entropy approximated using the analytical expression discussed in Section IV, for each training epoch outlined in Section V-A. With these data, we analyzed the convergence value of the guessing entropy n as a linear function of the convergence metric $\hat{n}_2$ by performing a least-squares linear regression. This approach yielded the results shown in Figure 4.

When approximating n as $n = \beta_0 + \beta_1\ \hat{n}_2$, we obtained $\beta_0 = 6.816E{-}5 \pm 5.712E{-}10$ and $\beta_1 = 0.9869 \pm 8.110E{-}7$. This indicates that the bias β_0 in the approximation is negligible in most cases. Nonetheless, as the theory in Section IV predicted, the estimated convergence should be greater than or equal to the actual value to which the guessing entropy converges. In fact, for most measurements, the convergence value obtained is either equal to or higher than the true value.

Furthermore, it is noteworthy that the range of n values where the most significant discrepancies between the estimated convergence and the actual convergence occur is within the interval $n \subset (70,\ 100)$). These values coincide with the initial improvements observed in the models' attack capabilities.

Given that $\hat{n}_2$ can approximate the values at which the guessing entropy functions converge, we can leverage it to examine the evolution of the models' attack capabilities throughout the training process. This analysis yields the results in Figure 5.

In Figure 5, we can observe how the metric informs us about the evolution of the models' ability to execute effective attacks throughout training and across different datasets. In the early training stages, the obtained values hover around 128, indicating that the models cannot achieve guessing entropy values lower than those produced by uniformly distributing the probability among all classes. After a certain number of epochs, the metric reveals an abrupt drop in the convergence value of the models, followed by stabilization. Notably, the number of epochs preceding this drop appears to depend more on the dataset used than on the type of initial layer in the model, rising as data desynchronization increases or

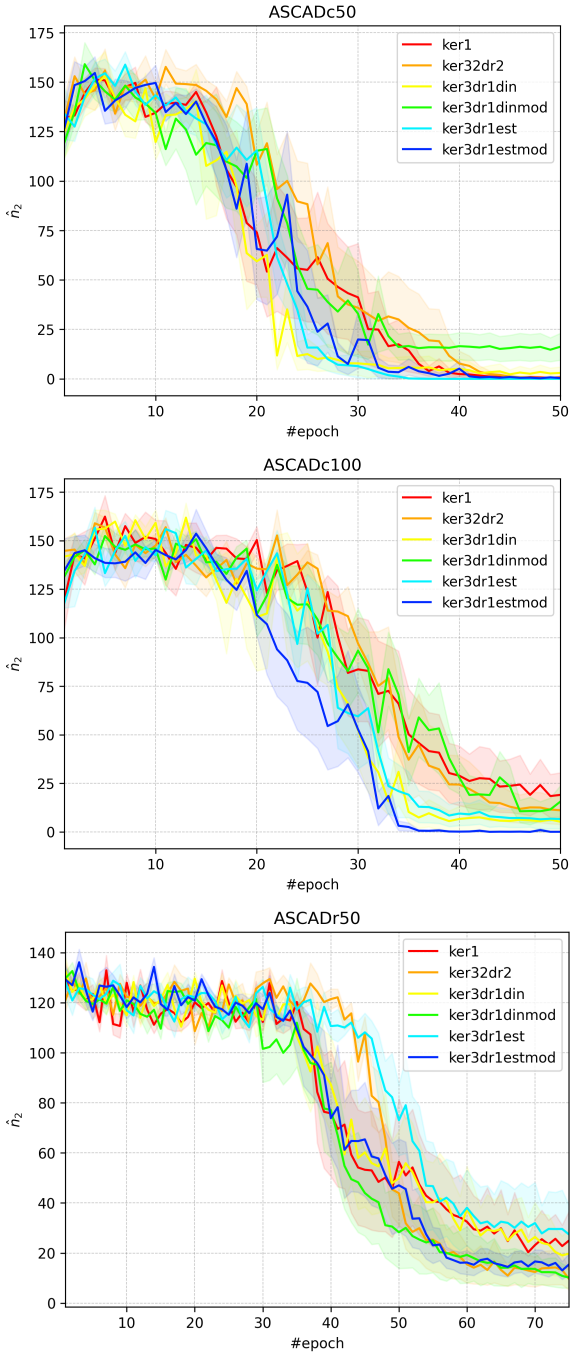


Fig. 5. Mean and standard deviation of $\hat{n}_2$ values for each epoch, along different datasets and models.

as the problem becomes more challenging, transitioning from ASCADf to ASCADr.

Similar to the observations in Section V-A, Figure 5 again demonstrates the regulatory effect of modulation. With the simpler data in ASCADf, modulation hinders performance, whereas in ASCADr, it proves beneficial by yielding lower $\hat{n}_2$ values compared to unmodulated models. Additionally, between the ker3dr1din and ker3dr1est models, the dynamic model generally exhibits lower convergence values during most of the training process. This trend holds even in AS-

CADf50, where the dynamic model outperforms the static model throughout the training - except in the final phase, where the static model eventually achieves a lower value, as shown in Figure 2. Once again, when considering the modulated versions, ker3dr1estmod emerges as the model with the lowest convergence values across most datasets. Furthermore, these modulated versions exhibit lower deviation values than the other models, indicating a higher predictability in their effectiveness.

VI. Conclusions

In this work, we introduced deformable convolutional models as a potential advancement over the current state-of-the-art models. We conducted a comparative analysis between two standardized convolutional models in side-channel analysis and their counterparts in which the initial layers were replaced by deformable convolutions, specifically static, dynamic, modulated static, and modulated dynamic types. The results indicate that the deformable variants achieve similar or superior performance in several cases compared to the reference models. Notably, among the deformable architectures, the modulated static model delivered the best performance, exhibiting some of the lowest guessing entropy values across multiple datasets and minimal deviation, demonstrating robust and consistent results across different training sessions.

Additionally, we computed the probability density function of the absolute deformation values of the models. This analysis served as an interpretability exercise, shedding light on the model's capacity to adapt filter shapes to the data. It was observed that dynamic models tend to perturb the filter shapes, while static models - despite being more likely to produce zero deformations - can also generate deformations with absolute values exceeding one, thus affecting significant alterations in the shape of the filters. Moreover, modulation enhanced the model's ability to produce larger deformations, which, when combined with the static approach, ultimately yielded the best-performing models.

A new performance evaluation metric, called convergence, has been designed to assess the efficacy of models in executing side-channel attacks. A theoretical analysis was conducted on the metric's deviation and bias relative to the convergence value of the guessing entropy, which is the target value that the metric seeks to approximate. An experimental evaluation of the metric's ability to approximate n was also performed. By executing a weighted linear regression between the two, we obtained $\beta_1 = 0.9869 \pm 8.110E-7$, while the bias β_0 was found to be several orders of magnitude lower. Furthermore, the metric was computed for all models trained on various datasets over all epochs, enabling a detailed study of the evolution of model performance throughout training.

Despite these results, some limitations were identified. Notably, the metric exhibits its maximum bias during the early stages of performance improvement in the attack capabilities. This introduces uncertainty by preventing an exact determination of the epoch at which the abrupt performance gains occur. Regarding the deformable models, our findings indicate that the ker3dr2estmod model was the most effective in executing side-channel attacks. However, it and the rest of the deformable models incurred computational costs up to 6 times higher than the baseline models, an important factor to consider.

In future work, we will leverage additional computational resources to investigate the performance effects of using larger filters in deformable networks. Additionally, we will explore the impact of implementing deformable convolutions in layers beyond the first one.

References

[1] N. I. of Standards, T. (NIST), M. J. Dworkin, M. S. Turan, and N. Mouha, "Advanced encryption standard (aes)," 2023-05-09 04:05:00 2023. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936594

[2] P. C. Kocher, "Timing attacks on implementations of diffie-hellman, rsa, dss, and other systems," in *Advances in Cryptology — CRYPTO '96*, N. Koblitz, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 104–113.

[3] P. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *Advances in Cryptology — CRYPTO' 99*, M. Wiener, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 388–397.

[4] E. Prouff, R. Strullu, R. Benadjila, E. Cagli, and C. Dumas, "Study of deep learning techniques for side-channel analysis and introduction to ASCAD database," Cryptology ePrint Archive, Paper 2018/053, 2018. [Online]. Available: https://eprint.iacr.org/2018/053

[5] G. Zaid, L. Bossuet, A. Habrard, and A. Venelli, "Methodology for efficient CNN architectures in profiling attacks," *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, vol. 2020, no. 1, pp. 1–36, 2020. [Online]. Available: https://doi.org/10.13154/tches.v2020.i1.1-36

[6] S. Paguada and I. Armendariz, "The forgotten hyperparameter:," in *Applied Cryptography and Network Security Workshops*, J. Zhou, M. Conti, C. M. Ahmed, M. H. Au, L. Batina, Z. Li, J. Lin, E. Losiouk, B. Luo, S. Majumdar, W. Meng, M. Ochoa, S. Picek, G. Portokalidis, C. Wang, and K. Zhang, Eds. Cham: Springer International Publishing, 2020, pp. 217–236.

[7] S. Hajra, S. Saha, M. Alam, and D. Mukhopadhyay, "TransNet: Shift invariant transformer network for side channel analysis," Cryptology ePrint Archive, Paper 2021/827, 2021. [Online]. Available: https://eprint.iacr.org/2021/827

[8] F.-X. Standaert, T. G. Malkin, and M. Yung, "A unified framework for the analysis of side-channel key recovery attacks," in *Advances in Cryptology - EUROCRYPT 2009*, A. Joux, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 443–461.

[9] G. Perin, L. Wu, and S. Picek, "The need for speed: A fast guessing entropy calculation for deep learning-based SCA," Cryptology ePrint Archive, Paper 2021/1592, 2021. [Online]. Available: https://eprint.iacr.org/2021/1592

[10] S. Paguada, L. Batina, I. Buhan, and I. Armendariz, "Being patient and persistent: Optimizing an early stopping strategy for deep learning in profiled attacks," *IEEE Transactions on Computers*, vol. PP, pp. 1–12, 01 2023.

[11] A. Ito, R. Ueno, and N. Homma, "Perceived information revisited ii: Information-theoretical analysis of deep-learning based side-channel attacks," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 2025, no. 1, p. 450–474, Dec. 2024. [Online]. Available: https://tches.iacr.org/index.php/TCHES/article/view/11936

[12] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[13] S. Chari, J. R. Rao, and P. Rohatgi, "Template attacks," in *Cryptographic Hardware and Embedded Systems - CHES 2002*, B. S. Kaliski, ç. K. Koç, and C. Paar, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 13–28.

[14] L. Wouters, V. Arribas, B. Gierlichs, and B. Preneel, "Revisiting a methodology for efficient cnn architectures in profiling attacks," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 2020, no. 3, p. 147–168, Jun. 2020. [Online]. Available: https://tches.iacr.org/index.php/TCHES/article/view/8586

[15] L. Xiao, H. Zhang, W. Chen, Y. Wang, and Y. Jin, "Transformable convolutional neural network for text classification," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 4496–4502. [Online]. Available: https://doi.org/10.24963/ijcai.2018/625

[16] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9300–9308.

Mejora del algoritmo de encriptado DRPE para CGH de fase mediante barajado aleatorio de píxeles

Alfonso Blesa Gascón
Dpto. de Ing. Electrónica y Comunicaciones
Universidad de Zaragoza
EUPT. C/ Atarazanas, 2. Teruel
ablesa@unizar.es

Francisco José Serón Arbeloa
Dpto. de Informática e Ing. de Stmas.
Universidad de Zaragoza
EINA, C/ María de Luna, 3. Zaragoza
seron@unizar.es

Diego Sanz Fuertes
Dpto. de Informática e Ing. de Stmas.
Universidad de Zaragoza
EUPT. C/ Atarazanas, 2. Teruel
diego.sanz@unizar.es

***Resumen*—El encriptado óptico de información mediante hologramas generados por computador (CGH) es un campo bien estudiado que se basa en la modificación del patrón de fase obtenido a partir del cálculo del holograma de una escena. Este artículo presenta una evolución de este algoritmo que añade permutaciones aleatorias de píxeles. Se describe el flujo de trabajo, que incluye la síntesis y el almacenamiento de un CGH a partir de una escena 3D, los algoritmos de encriptado y desencriptado, y la visualización tanto en el laboratorio como mediante simulación. Concretamente, el artículo se centra en las propuestas relacionadas con el encriptado de la fase, permitiendo el uso, si se desea, de una única clave en los procesos de permutación de píxeles y de modificación de fase para generar el CGH encriptado. El método propuesto permite el uso de una o dos claves, en función del nivel de seguridad deseado, y añade una nueva capa de complejidad frente a ataques de fuerza bruta gracias al barajado de píxeles.**

***Index Terms*—CGH, Encriptado Óptico, DRPE**

Tipo de contribución: *Investigación original*

I. Introducción

Una imagen es una de las formas más naturales para recibir una gran cantidad de información. Por ello la ciencia y la tecnología siempre han mostrado un especial interés en conocer y modelar todo lo relativo a la formación, modificación, almacenamiento y transmisión de imágenes. La óptica es la disciplina que da cuerpo a este conocimiento adquirido durante siglos y ofrece un modelo de la propagación de la información y de la energía a través de las ondas electromagnéticas en el espectro visible [1].

Pero ha sido en estas últimas décadas cuando las técnicas fotónicas y ópticas se han mostrado como excelentes herramientas para procesar la información contenida en una imagen, sobre todo cuando los modelos teóricos desarrollados anteriormente se han podido concretar en algoritmos ejecutables por un ordenador. Al mismo tiempo, ha sido necesario desarrollar sistemas y algoritmos que protejan esta información de accesos no autorizados o de manipulación de datos [2], [3], [4].

De todas estas tecnologías, los Hologramas Generados por Computador (CGH) se han mostrado como una opción viable para el almacenamiento masivo de información, típicamente relacionada con imágenes de escenas 2D y 3D. Para ello se requiere de una gran cantidad de recursos computacionales que, en ocasiones, superan las capacidades de los computadores actuales si se quieren sintetizar en tiempos aceptables.

De forma análoga, la información almacenada en un CGH se debe poder almacenar y transmitir de forma segura, por lo que, en paralelo, se han desarrollado técnicas de encriptado de información específicas para salvaguardar la privacidad e integridad de los CGH [5], [6].

Un método de encriptado óptico ampliamente utilizado para aplicaciones de imágenes es el *Double Random Phase Encoding* (DRPE) [7]. En este método, la imagen de entrada se representa mediante una matriz de intensidad de la luz, a la cual se le añade una matriz de fase aleatoria. A continuación, se aplica una segunda máscara de fase aleatoria en un sistema óptico de dos lentes. El resultado se considera la información encriptada. Una evolución de este trabajo es el encriptado óptico mediante holografía digital [8]. Esta técnica también se puede modelar mediante procedimientos algebraicos como las multiplicaciones vector-matriz [9].

El DRPE ha sido probado con éxito tanto con claves estáticas como dinámicas, pero esta técnica puede tener algunas debilidades frente a ataques reconocidos en la literatura: si el proceso es puramente lineal, existe la posibilidad de romper la clave [10]. Este método se ha mejorado con diversas propuestas, como la modificación de píxeles mediante funciones caóticas [11] o mediante técnicas de aprendizaje profundo [12].

Tradicionalmente, las máscaras de fase propuestas para el DRPE han tenido como objetivo garantizar la uniformidad tanto de la amplitud del holograma como de la distribución de fases del CGH encriptado para dificultar los accesos no autorizados. Para su implementación, se han utilizado diversos materiales fotosensibles (p. ej., medios fotorrefractivos) [13]. En la actualidad, gracias a la mejora de las tecnologías de moduladores espaciales de luz (SLM, *Spatial Light Modulator*), es posible presentar directamente a un haz de luz en el laboratorio los elementos difractivos calculados digitalmente (como los CGH y las máscaras de fase). Esto significa que tanto los datos a encriptar como las claves utilizadas en el proceso de encriptado pueden ser dinámicas.

En este artículo se presenta una evolución del método DRPE al que se le añade una permutación aleatoria de píxeles. En la sección II, se describe el flujo de trabajo completo, que comienza con la síntesis y el almacenamiento de un CGH a partir de una escena 3D, continúa con los algoritmos de encriptado y desencriptado, y concluye con su visualización final, tanto en el laboratorio como mediante simulación.

Este trabajo se centra específicamente en las propuestas relacionadas con el proceso de encriptado. La novedad de este enfoque radica en que este procedimiento permite, si se desea, utilizar una única clave tanto en el proceso de

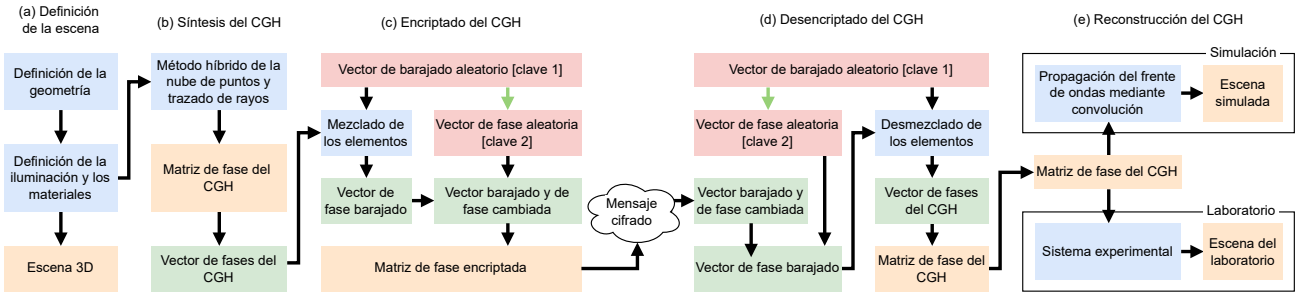


Fig. 1. Diagrama de flujo del método para generar un CGH a partir de una escena virtual, encriptado, desencriptado y visualización: (a) modelado de la escena, (b) síntesis del CGH mediante las técnicas de nube de puntos y trazado de rayos, (c) encriptado del CGH en dos pasos: barajado de píxeles y modificación aleatoria de fase, (d) desencriptado, (e) visualización tanto a través de simulación como en el laboratorio.

permutación de píxeles como para la modificación de fase, generando el CGH encriptado; o utilizar una clave distinta en cada uno de los procesos, aumentando el nivel de seguridad. En la sección III se valoran los resultados obtenidos. En la sección IV se presentan posibles líneas de investigación que podrían desarrollarse a partir de los resultados obtenidos en este trabajo. Finalmente, en la sección V se destacan las principales conclusiones de este trabajo.

II. De la síntesis de una escena 3D a su recepción segura: Descripción del proceso

El diagrama de flujo del proceso propuesto se muestra en la Fig. 1. En ella se muestran las diferentes etapas necesarias para enviar de forma segura una escena 3D: modelado de la escena (Fig. 1a), síntesis del CGH mediante las técnicas de nube de puntos y trazado de rayos (Fig. 1b), encriptado del CGH en dos pasos: barajado de píxeles y modificación aleatoria de fase (Fig. 1c), desencriptado (Fig. 1d) y visualización tanto a través de simulación como en el laboratorio (Fig. 1e). En azul, se muestran los algoritmos y procedimientos; en naranja, los resultados que se pueden visualizar (matrices de fase e imágenes); en verde, los vectores intermedios utilizados para manipular la información durante el proceso de encriptado: y en rojo, las claves utilizadas para encriptar o desencriptar. Las flechas verdes representan el uso de una clave única.

II-A. Escena y síntesis del CGH

Para mostrar el algoritmo de encriptado utilizado, se ha diseñado una escena 3D con objetos bien conocidos en el ámbito de la computación gráfica: un dragón de Stanford y una tetera de Utah contenidos en una caja de Cornell. Esta fase se corresponde con la Fig. 1a.

Mediante el uso híbrido de la técnica de la nube de puntos y la técnica de trazado de rayos [14], se calcula un CGH que se almacena en una matriz de $N \times M = 1920 \times 1080$ elementos (Fig. 1b). La información de la fase calculada se codifica en una imagen en formato PNG, que posteriormente se puede presentar a un SLM, en nuestro caso, el Pluto Holoeye [15] con un tamaño de píxel de $8\mu m$.

El cálculo del CGH ya incluye inherentemente la primera máscara de fase descrita en el algoritmo DRPE y el procedimiento presentado permite almacenar escenas 2D y 3D. Esta fase se corresponde con la Fig. 1b.

En la Fig. 2 se muestra la escena propuesta y en la Fig. 3 se muestra la fase del CGH calculado en formato PNG. Las

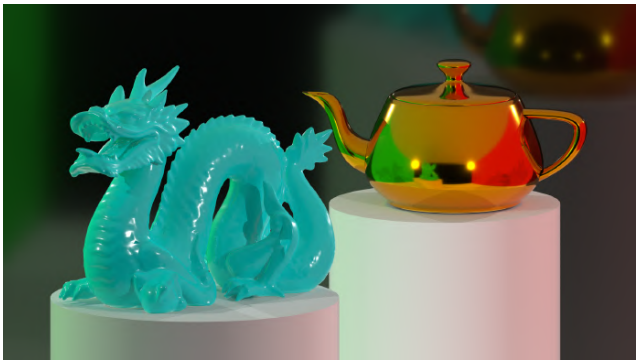


Fig. 2. Escena 3D inicial.

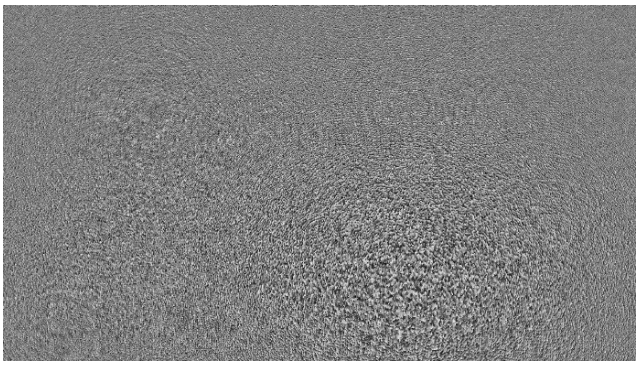


Fig. 3. Fase del CGH codificada en formato PNG para su posterior transmisión a un SLM.

dimensiones de la ventana a la escena coinciden con las del SLM ($8mm \times 15mm$). El proceso de síntesis de la escena y generación del CGH se puede encontrar con detalle en [16].

La modificación del tamaño del píxel o del SLM afecta directamente al número de elementos en la matriz de fase y, por ende, a la calidad del holograma generado. Cuantos más elementos haya, mayor será la dificultad de un ataque de fuerza bruta, aunque esto implicará un mayor coste computacional en la generación y recuperación del CGH.

II-B. Encriptado del CGH

En la Fig. 4 se muestra el diagrama de flujo de generación de claves y encriptado del CGH. El método propuesto implica aplicar el algoritmo de DRPE para la modificación de fase y el barajado de píxeles del CGH. Veremos que el barajado

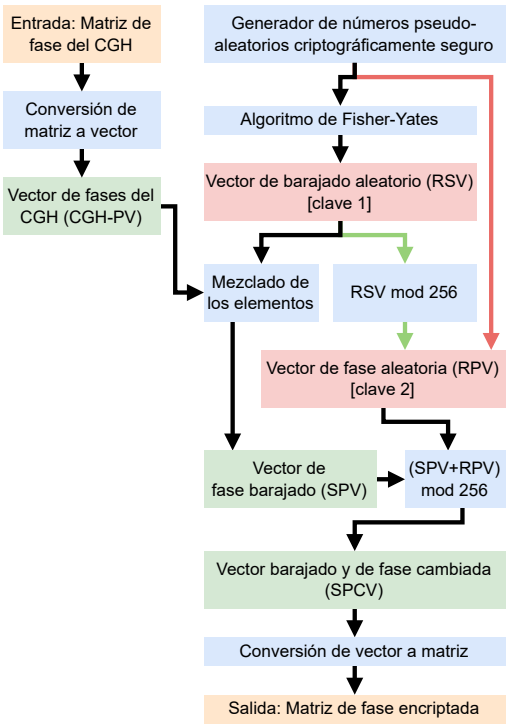


Fig. 4. Diagrama de flujo del proceso de encriptado. La flecha roja indica el diagrama de flujo para generar dos claves independientes.

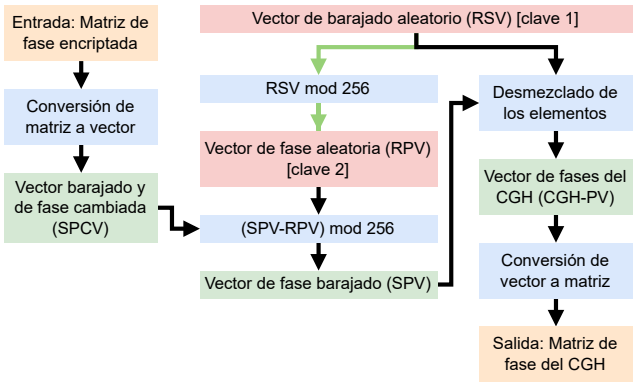


Fig. 5. Diagrama de flujo del proceso de desencriptado.

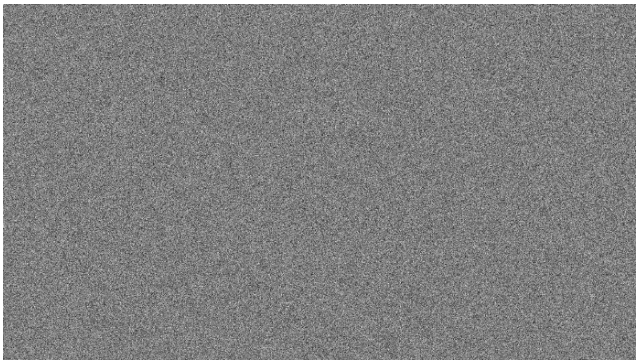


Fig. 6. Fase del CGH encriptada mediante DRPE y barajado de píxeles, en formato PNG. Este es el mensaje encriptado.

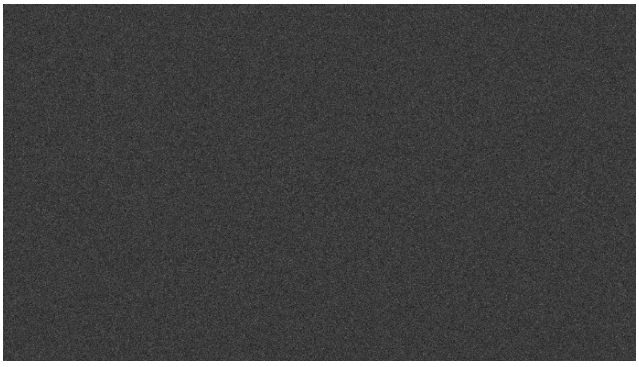


Fig. 7. Imagen obtenida al desencriptar el CGH con una clave incorrecta y simular su propagación mediante el *Angular Spectrum Method*.

de píxeles genera una explosión combinatoria difícil de atacar por fuerza bruta. Por otra parte, la modificación de fase añade una capa más de encriptado a la información almacenada en el CGH. Este proceso se corresponde con la Fig. 1c.

Para la elección de una clave que sea verdaderamente aleatoria se debe utilizar un generador de números pseudo-aleatorios que sea criptográficamente seguro (CSPRNG, *Cryptographically Secure Pseudo-Random Number Generator*). Estos generadores se pueden encontrar en librerías criptográficas como OpenSSL.

Estos números permiten alimentar el algoritmo de Fisher-Yates [17] para barajar un conjunto de elementos de manera aleatoria y segura. Este algoritmo consiste en recorrer los elementos de un vector en orden inverso y generar un índice aleatorio j para cada elemento i, de tal forma que $0 \leq j \leq i$, siendo i el índice del elemento que se modifica en cada momento. Finalmente se intercambian los elementos i y j de posición.

Esta información se guarda en el vector de barajado aleatorio (RSV, *Random Sorting Vector*) de tamaño $Z = M \times N$. El proceso de encriptado utiliza la matriz de fases del CGH guardada en un fichero PNG para convertirla en un vector CGH-PV de tamaño Z. Utilizando el RSV se cambian las posiciones de CGH-PV, obteniendo el vector de fase barajado (SPV, *Shuffled Phase Vector*).

En función del nivel de seguridad requerido, se puede obtener un vector de fases aleatorias (RPV, *Random Phase Vector*) mediante una operación de módulo sobre el RSV, ya que los valores de la fase se encuentran en el intervalo $[0\,..\,255]$:

$$\forall i \in [1\,..\,Z],\ RPV[i] = RSV[i]\ \%\ 256 \qquad (1)$$

o mediante un procedimiento independiente (siguiendo la flecha roja en la Fig. 4).

Este vector se suma al SPV para modificar el patrón de fases, obteniendo el vector barajado y de fase cambiada (SPCV, *Shuffled and Phase Changed Vector*):

$$\forall i \in [1\,..\,Z],\ SPCV[i] = (SPV[i] + RPV[i])\ \%\ 256 \qquad (2)$$

El SPCV se reordena para recuperar una matriz de tamaño $M \times N$ que será el CGH encriptado a enviar a un receptor.

El resultado obtenido al encriptar el CGH mostrado en la Fig. 3 se muestra en la Fig. 6. La imagen obtenida al no desencriptar, o desencriptar con una clave incorrecta, y visualizar el CGH se muestra en la Fig. 7.

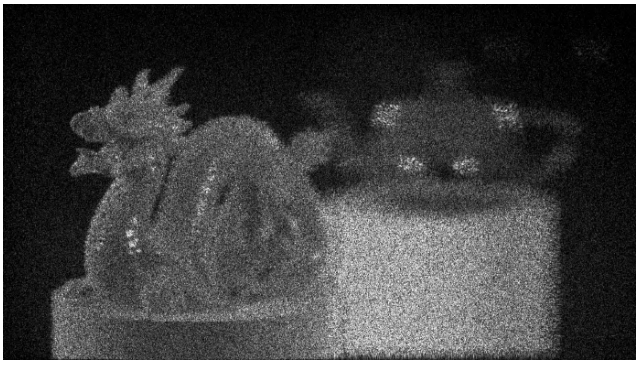


Fig. 8. Imagen obtenida al desencriptar el CGH con la clave correcta y simular su propagación mediante el *Angular Spectrum Method*.

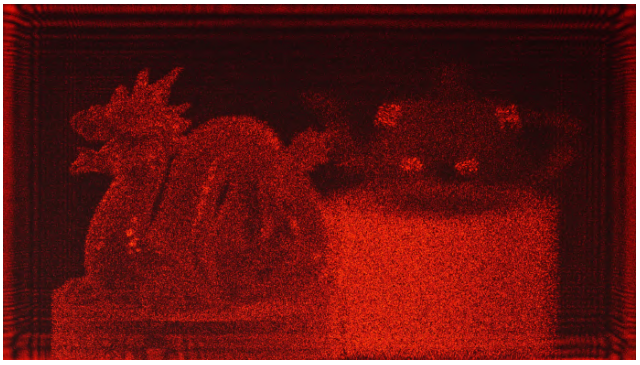


Fig. 9. Imagen obtenida al desencriptar el CGH con la clave correcta en el laboratorio.

II-C. Desencriptado del CGH

El desencriptado, mostrado en las Figs. 1d y 5, implica deshacer los pasos de la sección anterior.

Se comienza convirtiendo la matriz de fases del CGH encriptado a un vector, siendo éste el SPVC. Restar el RPV al SPVC permite recuperar el SPV:

$$\forall i \in [1\,..\,Z],\ SPV[i] = (SPVC[i] - RPV[i])\ \%\,256 \quad (3)$$

El orden original es restaurado utilizando los índices almacenados en el RSV para obtener el CGH-PV, que se convierte en una matriz de fases, obteniendo el CGH desencriptado.

Este proceso de encriptado y desencriptado garantiza que el CGH recuperado no pierde la información de la matriz de fases inicial.

II-D. Visualización del CGH

La escena original se puede recuperar mediante simulación, utilizando algoritmos de convolución como el *Angular Spectrum Method* [1], (Fig. 8) o en un laboratorio de óptica holográfica (Fig. 9). En ambos casos, se visualiza el comportamiento de un frente de ondas plano perturbado con la información de la fase del CGH en cada píxel, en un plano concreto que coincide con alguna posición de los objetos de la escena. Para obtener la imagen de la Fig. 9, se sitúa en este plano el sensor de una cámara fotográfica sin objetivo.

La calidad de la imagen final suele verse afectada por varios motivos: trabajar solo con la fase y codificarla con números enteros en el rango admitido, así como el rango dinámico o el tamaño de píxel del SLM (real o simulado). Siempre aparecen fenómenos de moteado interferencial (*speckle*) que no son atribuibles a la técnica de encriptado. Para mejorar la imagen final, se puede iterar el proceso generando varios CGH y aplicando técnicas de filtrado estadístico [18], [14].

III. Discusión de los resultados

El procedimiento propuesto es robusto frente a distintos tipos de ataques, ya que:

- **Almacenado en CGH**: La escena 3D se almacena en un CGH. De esta manera, se mitiga la vulnerabilidad al análisis estadístico mediante métodos como histogramas de intensidad (Fig. 10), que son incapaces de revelar patrones o facilitar la reconstrucción parcial de la imagen. Además, es resistente a transformaciones parciales, ya que la modificación de un subconjunto de píxeles en el CGH no compromete la integridad de la información almacenada.
- **Barajado de píxeles**: Como se demostrará a continuación, el barajado de píxeles puede ser robusto contra ataques de fuerza bruta cuando el número de permutaciones de píxeles posibles es suficientemente grande.
- **Modificación de fase aleatoria**: Esta es la base del DRPE. Añade una capa adicional de complejidad a la encriptación, ya que modifica la fase que cada píxel del CGH presenta a la reconstrucción del frente de ondas. También altera el histograma de fase del CGH encriptado, lo que dificulta aún más los ataques utilizando herramientas estadísticas.

III-A. Observaciones sobre el barajado de píxeles

Se parte de una matriz de $Z = N \times M$ números enteros, que pueden tener valores en el rango $[0\,..\,255]$; esto es, 256 valores diferentes.

Supongamos que cada valor de píxel aparece repetido una cantidad determinada R_i; esto es, $(R_1, R_2, R_3, \ldots, R_{256})$, donde R_i puede tomar un valor comprendido en el intervalo $[1\,..\,Z]$ con la restricción de que la suma de todos los valores R_i, será Z.

Los Z píxeles de la matriz, se extraen sin reemplazo y se colocan un vector de tamaño $N \times M$ píxeles. Las extracciones se supone que son estadísticamente independientes.

Comenzando con la extracción, recorriendo Z:

- Probabilidad de que el primer píxel escogido coincida con el primer píxel de la imagen: Supongamos que el primero tiene el valor j comprendido entre $[0\,..\,255]$, y aparece repetido R_j veces, entonces:

$$P_1 = \frac{R_j}{N \times M} \quad (4)$$

- Probabilidad de que la segunda extracción coincida con el segundo píxel de la imagen: Supongamos que el segundo tiene el valor k comprendido entre $[0\,..\,255]$, y aparece repetido R_k veces, entonces:

$$P_2 = \frac{R_k}{(N \times M) - 1} \quad (5)$$

pero también puede ocurrir que tenga un valor j, con lo que la probabilidad sería:

$$P_2 = \frac{R_j - 1}{(N \times M) - 1} \quad (6)$$

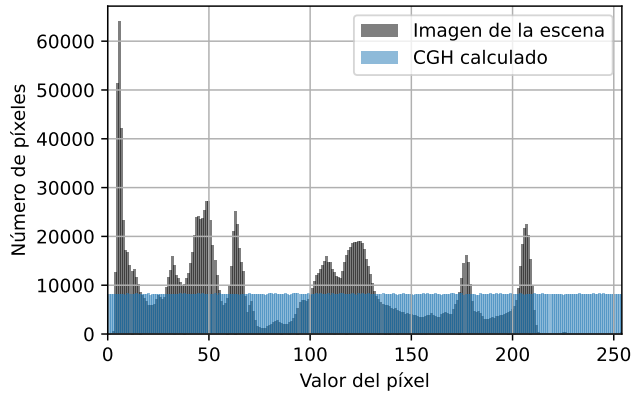


Fig. 10. Histograma de la luminancia de la escena 3D inicial (Fig. 2), en gris, y de las fases del CGH calculado (Fig. 3), en azul.

- Siguiendo este razonamiento, la probabilidad de que la extracción nº L coincida con el píxel nº L de la imagen. Supongamos que el pixel nº L tiene el valor m comprendido entre $[0\,..\,255]$, y aparece repetido R_m veces pero ha salido n veces anteriormente, entonces:

$$P_L = \frac{R_m - n}{(N \times M) - (L-1)} \tag{7}$$

Iterando este proceso, la probabilidad final de que la extracción resulte en la reconstrucción de la imagen se determina por la probabilidad de acertar con todos los píxeles del CGH; esto es:

$$P = \prod_{i=1}^{Z} P_i = \frac{\prod_{i=1}^{256} R_i!}{Z!} \tag{8}$$

La explosión combinatoria está determinada por el rango que los valores pueden tomar, así como por las dimensiones de la matriz. En nuestro caso, el rango es $[0\,..\,255]$ y la resolución es $Z = 1920 \times 1080$. Para realizar una estimación, sabemos que los histogramas de las fases codificadas en un fichero PNG tienden a ser prácticamente uniformes (Fig. 10). Los valores de frecuencia en estos histogramas tienden a cumplir la siguiente igualdad: $R_1 = R_2 = \cdots = R_{256} = Z/256$. Por tanto, podemos realizar la siguiente estimación:

$$P = \frac{(Z/256)!^{256}}{Z!} = \frac{8\,100!^{256}}{2\,073\,600!} \tag{9}$$

Los factoriales son números extraordinariamente grandes. Mediante la aproximación de Stirling [19] se puede hacer una estimación:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{10}$$

para conocer el tamaño de este número podemos calcular los dígitos que tiene:

$$\log_{10}(n!) \approx n\log_{10}(n) - n \tag{11}$$

esto da como aproximación:

$$P \approx \frac{(10^{23\,558})^{256}}{10^{11\,024\,760}} = \frac{10^{6\,030\,848}}{10^{11\,024\,760}} = \frac{1}{10^{4\,993\,912}} \tag{12}$$

El denominador tiene alrededor de casi cinco millones de dígitos, por lo tanto, la probabilidad de reconstrucción de la matriz $N \times M$ es prácticamente nula. Si se utilizara una máquina de procesamiento a la que le costará $10^{-12}s$ cada posibilidad, serían necesarios $10^{4\,993\,912} \times 10^{-12}s$ que es un periodo de tiempo muy largo en comparación con cualquier escala temporal humana o incluso geológica (el universo tiene una edad en la escala de $10^{17}s$).

Este método de barajado de píxeles de una escena en el que previamente se ha realizado una operación de transformación matemática (la síntesis de un CGH se puede asimilar a este operador), ya proporciona un encriptado bastante robusto frente a análisis estadísticos o ataques de fuerza bruta.

III-B. Observaciones sobre la máscara del DRPE

Como se ha descrito en la subsección II-B, se dispone de la clave, que es un vector en el que aparecen las permutaciones de los píxeles (RSV), y de un vector en el que se han almacenado las fases permutadas (SPV). Ahora se van a modificar dichos valores del siguiente modo: se va a construir otro vector que almacenará las modificaciones de los valores de la fase a partir del vector en el que aparecen las permutaciones de los píxeles (RPV). Para ello, como Z es un número mayor que el rango de fases codificado en el CGH, el procedimiento propuesto consiste en obtener el módulo 256 de cada componente del RSV (clave única) o generar otro vector a partir de otra clave para mayor seguridad (doble clave), y añadirlo a la fase que aparece en el SPV en la misma posición, lo que da como resultado el vector barajado y de fase cambiada (SPCV). De esta forma, para cada píxel del CGH se obtiene un nuevo valor de fase que se encuentra en el rango $[0\,..\,255]$. Un ataque por fuerza bruta supondría modificar las posiciones y los valores de los píxeles.

Siguiendo el razonamiento de la subsección anterior, suponemos que tenemos Z elementos de una matriz con valores en un rango entre $[0\,..\,255]$. Añadir una fase aleatoria implica que cada uno de estos Z elementos puede tener cualquier valor en el rango mencionado. La probabilidad de recuperar la fase original del píxel i del CGH es:

$$P_i = \frac{1}{256} \tag{13}$$

Como cada píxel es un caso independiente, la probabilidad de acertar con todas las fases correctas sería:

$$P = \prod_{i=1}^{Z} P_i = \frac{1}{256^Z} = \frac{1}{256^{(1920\times1080)}} \tag{14}$$

La probabilidad de encontrar la relación de fases también es muy pequeña, ya que el denominador tiene casi 5 millones de dígitos, lo que añade una capa más de seguridad al encriptado del CGH, rompiendo cualquier patrón que pudiera haber en el histograma de fase.

III-C. Cálculo de la entropía

La entropía de una imagen en escala de grises viene dada por la siguiente expresión:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{15}$$

Donde X es la imagen, n es el número de píxeles, x_i es el valor del píxel i-ésimo, y p es la probabilidad de que el píxel tenga ese valor.

Tabla I
ENTROPÍA DE LAS DIFERENTES ETAPAS DEL PROCESO

Descripción	Figura	Entropía H(X)
Escena original	Fig. 2	7.286206
CGH	Fig. 3	7.994273
CGH encriptado	Fig. 6	7.994281
Imagen recuperada (simulación)	Fig. 8	5.191329
Imagen recuperada (lab)	Fig. 9	6.013095

En la Tabla I se presentan los valores de entropía obtenidos para las imágenes objeto de este trabajo, tras convertirlas a escala de grises.

Estos resultados confirman que la entropía de la información almacenada en el CGH es superior a la de las imágenes y no varía durante el proceso de encriptado. La diferencia de entropía entre la imagen original y las imágenes recuperadas se debe a las limitaciones físicas del sistema, como los efectos de difracción, el *speckle* y el tamaño de los píxeles.

Los valores de entropía del CGH y el CGH encriptado son prácticamente idénticos, ya que los histogramas de ambos presentan una distribución uniforme.

IV. TRABAJO FUTURO

Las claves utilizadas en el proceso de encriptado de información son los vectores de barajado aleatorio y el vector de fase aleatoria. Dado que la seguridad es del orden de 2^{-K}, donde K es el número de bits de la clave, esta propuesta puede alcanzar niveles de seguridad muy altos. En este trabajo se propone el uso de un generador de números aleatorios para generar estos vectores, pero queda como trabajo futuro definir el método para generarlos de forma más precisa.

V. CONCLUSIONES

Codificar información mediante CGH y encriptarla utilizando fases aleatorias se ha demostrado como una técnica robusta para encriptar información en el pasado. No obstante, el ataque por fuerza bruta acorta el tiempo de éxito a medida que se aumenta la capacidad de cómputo disponible para encontrar la clave de desencriptado.

Presentamos una evolución del algoritmo DRPE a la que se añade una permutación aleatoria de píxeles. Esta propuesta permite utilizar una única clave para generar el CGH encriptado o utilizar dos claves independientes, una para la permutación y otra para la síntesis de la máscara de fases aleatoria. El método de cálculo de la segunda clave puede ser más complejo que un simple módulo, dependiendo del nivel de seguridad deseado.

Todo el proceso se puede simular en un ordenador o bien mostrar en un laboratorio. Las posibilidades de permutación añaden una nueva capa de complejidad frente a ataques de fuerza bruta.

Encriptar la información en un CGH hace que sea robusta frente a modificaciones puntuales del mismo, ya que la naturaleza de los CGH implica que cada píxel contiene información sobre toda la escena. La validez de esta información depende de la posición del píxel. El DRPE hace que la distribución del histograma de fases sea más robusta frente a ataques estadísticos. El barajado de píxeles dificulta la posibilidad de desencriptado parcial, además de aumentar significativamente las posibilidades combinatorias a explorar, haciendo inviable un ataque por fuerza bruta.

AGRADECIMIENTOS

Esta iniciativa está financiada por la Unión Europea a través del "Plan de Recuperación, Transformación y Resiliencia (*Next Generation*), proyecto estratégico C077/23 del Instituto Nacional de Ciberseguridad (INCIBE)".

REFERENCIAS

[1] J. W. Goodman, "Introduction to fourier optics," *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2017*, vol. 1, 2017.

[2] W. Chen, B. Javidi, and X. Chen, "Advances in optical security systems," *Adv. Opt. Photon.*, vol. 6, no. 2, pp. 120–155, Jun 2014. [Online]. Available: https://opg.optica.org/aop/abstract.cfm?URI=aop-6-2-120

[3] N. K. Nishchal, *Optical Cryptosystems*, ser. 2053-2563. IOP Publishing, 2019. [Online]. Available: https://doi.org/10.1088/978-0-7503-2220-1

[4] M. Singh and A. K. Singh, "A comprehensive survey on encryption techniques for digital images," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 155–11 187, mar 2023. [Online]. Available: https://doi.org/10.1007/S11042-022-12791-6

[5] B. Javidi, A. Carnicer, M. Yamaguchi, T. Nomura, E. Pérez-Cabré, M. S. Millán, N. K. Nishchal, R. Torroba, J. F. Barrera, W. He, X. Peng, A. Stern, Y. Rivenson, A. Alfalou, C. Brosseau, C. Guo, J. T. Sheridan, G. Situ, M. Naruse, T. Matsumoto, I. Juvells, E. Tajahuerce, J. Lancis, W. Chen, X. Chen, P. W. H. Pinkse, A. P. Mosk, and A. Markman, "Roadmap on optical security," *Journal of Optics*, vol. 18, no. 8, p. 083001, jul 2016. [Online]. Available: https://doi.org/10.1088/2040-8978/18/8/083001

[6] X. Wang and M. Zhao, "An image encryption algorithm based on hyperchaotic system and DNA coding," *Optics & Laser Technology*, vol. 143, p. 107316, nov 2021. [Online]. Available: https://doi.org/10.1016/J.OPTLASTEC.2021.107316

[7] P. Refregier and B. Javidi, "Optical image encryption based on input plane and Fourier plane random encoding," *Optics Letters, Vol. 20, Issue 7, pp. 767-769*, vol. 20, no. 7, pp. 767–769, apr 1995. [Online]. Available: https://doi.org/10.1364/OL.20.000767

[8] E. Tajahuerce and B. Javidi, "Encrypting three-dimensional information with digital holography," *Applied Optics*, vol. 39, no. 35, pp. 6595–6601, dec 2000. [Online]. Available: https://doi.org/10.1364/AO.39.006595

[9] M. Takeda, K. Nakano, H. Suzuki, and M. Yamaguchi, "Encrypted imaging based on algebraic implementation of double random phase encoding," *Applied Optics, Vol. 53, Issue 14, pp. 2956-2963*, vol. 53, no. 14, pp. 2956–2963, may 2014. [Online]. Available: https://doi.org/10.1364/AO.53.002956

[10] A. Carnicer, M. Montes-Usategui, S. Arcos, and I. Juvells, "Vulnerability to chosen-cyphertext attacks of optical encryption schemes based on double random phase keys," *Opt. Lett.*, vol. 30, no. 13, pp. 1644–1646, Jul 2005. [Online]. Available: https://doi.org/10.1364/OL.30.001644

[11] I. A. Hamadi, R. K. Jamal, and S. K. Mousa, "Image encryption based on computer generated hologram and Rossler chaotic system," *Optical and Quantum Electronics*, vol. 54, no. 1, pp. 1–12, jan 2022. [Online]. Available: https://doi.org/10.1007/S11082-021-03406-9

[12] T. Hu, Y. Ying, X. Sun, and W. Jin, "The recovery scheme of computer-generated holography encryption–hiding images based on deep learning," *Optics Communications*, vol. 529, p. 129100, feb 2023. [Online]. Available: https://doi.org/10.1016/J.OPTCOM.2022.129100

[13] T. Nomura and B. Javidi, "Optical encryption based on the input phase mask designed for the space bandwidth of the optical system," in *Optical Information Systems III*, B. Javidi and D. Psaltis, Eds., vol. 5908, International Society for Optics and Photonics. SPIE, 2005, p. 59080B. [Online]. Available: https://doi.org/10.1117/12.618271

[14] D. Blinder, T. Birnbaum, T. Ito, and T. Shimobaba, "The state-of-the-art in computer generated holography for 3D display," *Light: Advanced Manufacturing*, vol. 3, no. 3, pp. 572–600, 2022. [Online]. Available: https://doi.org/10.37188/lam.2022.035

[15] "PLUTO-2 Phase Only Spatial Light Modulator (Reflective) — HOLOEYE Photonics AG." [Online]. Available: https://holoeye.com/spatial-light-modulators/slm-pluto-phase-only/

[16] J. A. Magallón, A. Blesa, and F. J. Serón, "Monte–carlo techniques applied to cgh generation processes and their impact on the image quality obtained," *Engineering Reports*, vol. 7, no. 1, p. e13109, 2025. [Online]. Available: https://doi.org/10.1002/eng2.13109

[17] R. Durstenfeld, "Algorithm 235: Random permutation," *Communications of the ACM*, vol. 7, no. 7, p. 420, 1964. [Online]. Available: https://doi.org/10.1145/364520.364540

[18] V. Bianco, P. Memmolo, M. Leo, S. Montresor, C. Distante, M. Paturzo, P. Picart, B. Javidi, and P. Ferraro, "Strategies for reducing speckle noise in digital holography," *Light: Science and Applications*, vol. 7, no. 1, 2018. [Online]. Available: https://doi.org/10.1038/s41377-018-0050-9

[19] D. E. Knuth, *The Art of Computer Programming, Volume 1: Fundamental Algorithms*, 3rd ed. Addison-Wesley, 1997, includes discussion on Stirling's approximation in combinatorial analysis. [Online]. Available: https://dl.acm.org/doi/book/10.5555/260999

QZKP Basada en Codificación Conjugada para Autenticación de Usuarios

Jorge Garcia-Diaz
Universidad de La Laguna
Tenerife, Spain
jgarcidi@ull.edu

Daniel Escanez-Exposito
Universidad de La Laguna
Tenerife, Spain
jescanez@ull.edu.es

Pino Caballero-Gil
Universidad de La Laguna
Tenerife, Spain
pcaballe@ull.edu.es

Jezabel Molina-Gil
Universidad de La Laguna
Tenerife, Spain
jmmolina@ull.edu.es

***Resumen*—Las Pruebas de Conocimiento Nulo (ZKPs, Zero-Knowledge Proofs) se utilizan ampliamente para la autenticación de usuarios, basándose tradicionalmente en problemas computacionalmente difíciles. Sin embargo, los recientes avances en computación cuántica suponen un peligro para muchos de esos esquemas, ya que podrían permitir la resolución eficiente de diversos problemas. Este trabajo propone una Prueba Cuántica de Conocimiento Nulo (QZKP, Quantum Zero-Knowledge Proof) para la autenticación de usuarios, que utiliza la codificación conjugada para aprovechar las relaciones entre los estados cuánticos que son opuestos en términos de las bases de medición. El protocolo garantiza la seguridad mediante la aleatoriedad inherente a las proyecciones de estados cuánticos en bases desalineadas, producida por la incertidumbre introducida cuando se miden los estados cuánticos en bases que no son ortogonales. Se proporciona un estudio de su completitud, solidez y conocimiento nulo, propiedades validadas además mediante simulaciones con el simulador cuántico Qiskit. Los resultados demuestran, además, que el protocolo mantiene su efectividad con recursos cuánticos limitados, lo que lo convierte en una alternativa viable para la autenticación cuántica segura.**

***Index Terms*—Criptografía Cuántica, Prueba Cuántica de Conocimiento Nulo, Protocolo de Autenticación, Prueba de Conocimiento Nulo, Distribución Cuántica de Claves**

Tipo de contribución: *Investigación original*

I. Introducción

Una de las herramientas más útiles para la autenticación de usuarios son las Pruebas de Conocimiento Nulo (ZKPs, Zero-Knowledge Proofs). Tradicionalmente, estos protocolos criptográficos se basan en problemas matemáticos que se consideran intratables desde el punto de vista computacional, comúnmente clasificados como problemas NP-completos. Sin embargo, con el rápido avance de la computación cuántica, algunos problemas que antes se consideraban intratables, han visto reducida su dificultad computacional gracias a algoritmos cuánticos como el algoritmo de búsqueda de Grover [1], mientras que otros han pasado a ser eficientemente resolubles, como la factorización de enteros mediante el algoritmo de Shor [2]. Aunque los esfuerzos de investigación se han dirigido a aprovechar las ventajas computacionales cuánticas sobre los sistemas clásicos [3], sigue habiendo una notable escasez de trabajos académicos que exploren la implementación de ZKPs dentro del paradigma cuántico y los marcos de Distribución Cuántica de Claves (QKD, Quantum Key Distribution) [4]. Específicamente, poca literatura aborda las ZKPs que operan independientemente de suposiciones matemáticas computacionalmente difíciles, abriendo una intrigante vía para futuras investigaciones en este campo.

Este trabajo propone una Prueba Cuántica de Conocimiento Nulo (QZKP, Quantum Zero-Knowledge Proof) para la autenticación de usuarios, que utiliza codificación conjugada [5], como el protocolo BB84 [6]. En la QZKP propuesta, se parte de la hipótesis de que probadora y verificador comparten una clave secreta, la cual debe haber sido distribuida previamente mediante alguna QKD conocida [7]. La seguridad de la propuesta se basa fundamentalmente en la aleatoriedad inherente del resultado cuando un estado cuántico dado se somete a proyección en una base diferente a la elegida para codificar. Esta aleatoriedad garantiza que cualquier desviación de la base de proyección prevista comprometa la capacidad de extraer información significativa, salvaguardando así el conocimiento nulo del protocolo.

En este trabajo se proporciona un estudio de la seguridad del protocolo, basada en las distribuciones de probabilidad de los probadores honestos y deshonestos. Este análisis permite evaluar correctamente las propiedades de completitud, solidez y conocimiento nulo del protocolo. Además, se corrobora este marco teórico mediante una implementación basada en simulaciones utilizando el simulador cuántico de IBM, Qiskit. Este protocolo mantiene su seguridad frente a probadores y verificadores deshonestos, garantizando una protección robusta frente a intentos adversarios de comprometer su integridad. Además, como demuestran los resultados, el protocolo no sólo mantiene sus garantías de seguridad, sino que también sigue siendo práctico y viable en condiciones reales. El protocolo propuesto ofrece una flexibilidad significativa en la implementación de cada iteración, permitiendo estrategias de ejecución adaptables. Esta adaptabilidad resulta especialmente ventajosa en escenarios en los que los recursos cuánticos están restringidos, como canales y hardware cuánticos limitados, mejorando así su aplicabilidad en diversos entornos de comunicación cuántica.

Este artículo se estructura como sigue. Una sección introductoria de preliminares sienta las bases del trabajo. La segunda sección aborda el funcionamiento del protocolo propuesto. La tercera sección describe un análisis de la seguridad tanto para partes honestas como deshonestas, que es validado en la cuarta sección mediante simulaciones realizadas. Se cierra el artículo con algunas conclusiones y trabajos futuros.

II. Preliminares

En esta sección se introducen las bases para desarrollar este trabajo. Primero se introducen las propiedades que caracterizan a las ZKPs y luego se dan las bases de computación e información cuántica necesarias.

ISBN:978-84-10169-61-6

II-A. Pruebas de Conocimiento Nulo

Las pruebas de conocimiento nulo permiten a una probadora convencer a un verificador de que posee cierto conocimiento sin revelar ninguna información adicional sobre dicho conocimiento. Para que una ZKP sea válida, debe cumplir tres propiedades esenciales; completitud, solidez y conocimiento nulo [8].

- Completitud: Si la afirmación es verdadera y ambas partes siguen el protocolo correctamente, entonces el verificador siempre aceptará la prueba. Es decir, si la probadora tiene la información correcta, la prueba será válida con una alta probabilidad.
- Solidez: Si la afirmación es falsa, ninguna probadora (incluso si actúa maliciosamente) podrá convencer al verificador de que la afirmación es verdadera, salvo con una probabilidad extremadamente pequeña. Esto garantiza que el verificador no será engañado por una afirmación falsa.
- Conocimiento Nulo: Si la afirmación es verdadera, el verificador no obtiene ninguna información adicional sobre la afirmación más allá de su veracidad. La prueba no revela nada sobre el secreto en sí, solo que la probadora lo conoce.

II-B. Computación e Información Cuántica

En la computación clásica, los bits son la unidad básica de información. Estos bits pueden estar en un estado 0 o 1. Por otra parte, en la computación cuántica, se utilizan los cúbits como unidad mínima de información, los cuales pueden estar en un estado de superposición de estos dos estados. Se denominan a los kets $|0\rangle$ y $|1\rangle$ a los estados básicos para un sistema de un sólo cúbit [9].

Vectorialmente, se entienden los estados básicos $|0\rangle$ y $|1\rangle$ como los vectores:

$$|0\rangle = (1,0)^T = \begin{pmatrix}1\\0\end{pmatrix} \text{ y } |1\rangle = (0,1)^T = \begin{pmatrix}0\\1\end{pmatrix}.$$

Definición 1 (Vector de Estado de un Cúbit): Un vector de estado en superposición de un solo cúbit es una combinación lineal:

$$\alpha_0 \begin{pmatrix}1\\0\end{pmatrix} + \alpha_1 \begin{pmatrix}0\\1\end{pmatrix}, \text{ donde } \alpha_0, \alpha_1 \in \mathbb{C} \text{ y } ||\alpha_0||^2 + ||\alpha_1||^2 = 1.$$

Al observar (medir) el cúbit en el estado $\alpha_0 |0\rangle + \alpha_1 |1\rangle$ con respecto a la base $\{|0\rangle, |1\rangle\}$, este colapsará al estado básico $|0\rangle$ con probabilidad $||\alpha_0||^2$ o al estado $|1\rangle$ con probabilidad $||\alpha_1||^2$.

Hay dos estados que son bastante conocidos y utilizados. Aunque lo más común es que haya números dentro del ket, también es posible que contenga etiquetas, como es por ejemplo el caso de los estados $|+\rangle$ y $|-\rangle$, que forman una base de $\mathbb{C}^2$:

$$|+\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}} \quad \wedge \quad |-\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}}$$

Definición 2 (Producto de Kronecker): Sean $A, B \in \mathcal{M}(\mathbb{C})$, es decir, dos matrices cualesquiera con coeficientes complejos. Se define el producto de Kronecker de la siguiente manera:

$$A \otimes B = \begin{pmatrix} a_{11} \cdot B & \cdots & a_{1n} \cdot B \\ \vdots & \ddots & \vdots \\ a_{m1} \cdot B & \cdots & a_{mn} \cdot B \end{pmatrix}$$

Ya definido el producto de Kronecker, se puede interpretar un sistema de varios cúbits. Para n cúbits se tienen 2^n estados básicos:

$$|\underbrace{0\ldots0}_{n}\rangle, \ |\underbrace{0\ldots1}_{n}\rangle, \ \ldots, \ |\underbrace{1\ldots1}_{n}\rangle \in \mathbb{C}^{2^n}$$

Estos estados básicos forman una base ortonormal de $\mathbb{C}^{2^n}$. Matemáticamente, el estado $|a_1 \ldots a_n\rangle$ se interpreta como:

$$|a_1 \ldots a_n\rangle = \bigotimes_{i=1}^{n} |a_i\rangle, \tag{1}$$

donde $\otimes$ es el producto de Kronecker. Es importante mantener el orden de los $|a_i\rangle$ pues este producto no es conmutativo.

Definición 3 (Vector de Estado de n cúbits): Para varios cúbits, los estados de superposición son combinaciones lineales de la forma:

$$\alpha_0 |0\ldots0\rangle + \cdots + \alpha_{2^n-1} |1\ldots1\rangle, \ \alpha_i \in \mathbb{C} : \sum_{i=0}^{2^n-1} ||\alpha_i||^2 = 1$$

Definición 4 (Puerta Cuántica): Una puerta cuántica es un operador que actúa sobre vectores de estado de n cúbits y que viene caracterizado por una matriz unitaria.

Una de las puertas más útiles es la puerta H, la cual tiene un gran peso en este trabajo. Esta puerta actúa sobre los estados de la base computacional de la forma $H \cdot |0\rangle = |+\rangle$ y $H \cdot |1\rangle = |-\rangle$, además cumple la propiedad $H^2 = I$.

Otras puertas con gran impacto en este trabajo son las puertas X y Z. La puerta X actúa linealmente como una puerta lógica NOT sobre los estados $|0\rangle$ y $|1\rangle$, mientras que la Z actúa también de esta manera pero sobre los estados $|+\rangle$ y $|-\rangle$.

En este trabajo se utiliza el lenguaje algebraico descrito en [10], para facilitar la explicación del funcionamiento de la QZKP propuesta. Los siguientes conceptos serán de ayuda mas adelante para entender el funcionamiento del protocolo.

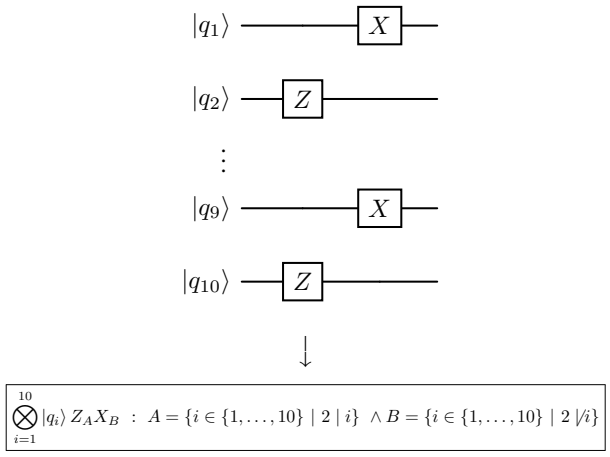
Notación 1: Sea A una puerta cuántica con el cúbit k como único cúbit objetivo sobre un registro cuántico de n cúbits y $k \in \{1, \ldots, n\}$. La acción de A sobre el estado $|a\rangle$ se expresa como:

$$|a\rangle \, A_k$$

Notación 2: Sea $R \subseteq \{1, \ldots, n\}$ y A una puerta cuántica cualquiera con un solo cúbit objetivo que actúe sobre el conjunto R de un registro cuántico de n cúbits $|a\rangle$. La acción de estas puertas sobre el estado $|a\rangle$ se denotará como se muestra a continuación:

$$|a\rangle \, A_R = |a\rangle \prod_{k \in R} A_k$$

Como ejemplo, se propone un circuito cuántico que implementa puertas Z para los qubits pares y puertas X para los qubits impares sobre un registro de 10 cúbits.



II-C. Codificación Conjugada

En general, siendo $A = \{|\gamma_1\rangle, |\gamma_2\rangle\}$ una base normal de $\mathbb{C}^2$, cualquier estado de la forma $|\psi\rangle = \beta_1 |\gamma_1\rangle + \beta_2 |\gamma_2\rangle$ se puede medir con respecto a la base A, con probabilidades $||\beta_1||^2$ y $||\beta_2||^2$ respectivamente.

Por ejemplo, si se considera la base $A = \{|+\rangle, |-\rangle\}$ y el estado:

$$\begin{aligned}\alpha_0 |0\rangle + \alpha_1 |1\rangle &= \alpha_0 \frac{|+\rangle + |-\rangle}{\sqrt{2}} + \alpha_1 \frac{|+\rangle - |-\rangle}{\sqrt{2}} \\ &= \frac{\alpha_0 + \alpha_1}{\sqrt{2}} |+\rangle + \frac{\alpha_0 - \alpha_1}{\sqrt{2}} |-\rangle\end{aligned}$$

efectuando la medición asociada a la base A al estado anterior, colapsará al estado $|+\rangle$ con probabilidad $||\frac{\alpha_0+\alpha_1}{\sqrt{2}}||^2$ o al estado $|-\rangle$ con probabilidad $||\frac{\alpha_0-\alpha_1}{\sqrt{2}}||^2$.

Es por esto que al medir tanto el cúbit $|0\rangle$ como $|1\rangle$ con respecto a la base $\{|+\rangle, |-\rangle\}$, la cual se suele llamar base de Hadamard, se obtienen tanto el estado $|+\rangle$ como el estado $|-\rangle$ con probabilidad $\frac{1}{2}$.

III. QZKP Propuesta

El protocolo propuesto en este trabajo es una QZKP en la que se presupone que tanto la probadora como el verificador comparten previamente un secreto.

III-A. Requisitos

La probadora (Alice) y el verificador (Bob) comparten un secreto formado por una secuencia de bits aleatorios $a \in \{0,1\}^n$, junto con una secuencia de bases aleatorias $b \in \{C, H\}^n$, donde C corresponde a la base Computacional y H a la base de Hadamard. Nótese que para dicha compartición pueden utilizar una QKD.

III-B. Ejecución del Protocolo

El primer paso es la inicialización mediante un *handshake* clásico que permite a Bob conocer la clave pública de Alice.

El segundo paso es la generación por parte del verificador del reto asociado al secreto compartido a, b. Para ello Bob genera un estado cuántico $|\psi\rangle$ utilizando las bases cuánticas secretas definidas por b para codificar los bits secretos a. Considerando que los bits 0 y 1 de b están relacionados, respectivamente, con la base de Hadamard y la base computacional y usando la notación algebraica introducida en [10], el estado cuántico resultante de este primer paso es $|\psi\rangle = |a\rangle X_{\{i|b_i=1\}}$. A continuación, Bob genera una cadena de bits aleatoria $c \in \{0,1\}^n$, y para cada $c_i = 1$ aplica una puerta X si $b_i = 0$ o Z si $b_i = 1$ en el i-ésimo cúbit del estado $|\psi\rangle$. Matemáticamente, esto se puede expresar como $|\psi\rangle X_{\Omega_0} Z_{\Omega_1}$, $\Omega_k = \{i \mid c_i = 1 \wedge b_i = k\}$. Bob envía entonces este *estado desafío* a Alice. El objetivo de Alice es proporcionar una estimación suficientemente precisa c' de c. Nótese que aplicar estas puertas tiene el mismo efecto que codificar la cadena $a \oplus c$ con respecto a las bases de b.

El tercer paso es la generación de la respuesta por parte de la probadora. Una vez que Alice recibe el estado $|\psi\rangle X_{\Omega_0} Z_{\Omega_1}$ ejecuta *modificaciones de conocimiento nulo*. Para ello, genera una cadena aleatoria de bits $d \in \{0,1\}^n$ y aplica la puerta Hadamard para cúbit donde $d_i = 1$ para obtener el estado $|\psi\rangle X_{\Omega_0} Z_{\Omega_1} H_\Gamma$, donde $\Gamma = \{i \mid d_i = 1\}$. Estas modificaciones aleatorias aseguran la propiedad de conocimiento nulo del protocolo aleatorizando la estimación de Alice c' si el *estado desafío* no se prepara honestamente, evitando así que verificadores deshonestos o terceras partes no obtengan información de las claves secretas. Después de las *modificaciones de conocimiento nulo*, Alice mide el estado resultante utilizando la base secreta b y obtiene una cadena de bits a'. A continuación, genera inteligentemente el *cadena de estimación* c' utilizando la información de Γ, a, y a'. Nótese que si $i \notin \Gamma$ Alice obtendrá como resultado de la medición el bit a_i original si $c_i = 0$ o $a_i \oplus 1$ si $c_i = 1$. En el caso de que $i \in \Gamma$ el resultado de la medición de Alice será totalmente aleatorio.

El cuarto paso es la verificación de la respuesta por parte del verificador. Para ello, al recibir c', Bob cuenta el número de coincidencias entre c y c', matemáticamente esto se puede representar como $n - w(c \oplus c')$, donde w representa el peso de Hamming. Si ambas partes son honestas, se espera que Alice recupere aproximadamente $75\,\%$ de los bits de c correctamente, mientras que si Alice es deshonesta, se espera que recupere sólo $50\,\%$ de c. Es por ello que se debe decidir un umbral de precisión para rechazar o aceptar al probador. La precisión de una probadora honesta sigue una distribución binomial $B(n, \frac{3}{4})$, mientras que una probadora deshonesta sigue una distribución binomial $B(n, \frac{1}{2})$. Esto hace que la seguridad del protocolo aumente al aumentar n.Además, no hay distinción probabilística entre ejecutar una única iteración para un tamaño de clave n o realizar k iteraciones sobre una clave secreta de longitud $\frac{k}{n}$ y tomar la precisión media de todas las iteraciones, lo que hace que ambas opciones sean igual de seguras excepto para el evidente ataque de fuerza bruta para tamaños de clave secreta menores.

Cabe mencionar también que el protocolo usa una prueba de n bits para mostrar el conocimiento de una clave de $2n$ bits. Del mismo modo el verificador solo debe comprobar la coincidencia de dos secuencias de bits de longitud n en vez de dos secuencias de longitud $2n$.

III-C. Generación de la prueba c'

Hay dos posibilidades cuando se mide cada cúbit: $a'_i = a_i$ o $a'_i \neq a_i$. Si $a_i \neq a'_i$, entonces i debe pertenecer al conjunto Γ o bien $c_i = 1$. En base a esto, se hacen las siguientes estimaciones:

- Si $i \in \Gamma$, entonces c_i se estima que es 0 ($c'_i = 0$).
- Si $i \notin \Gamma$, entonces c_i se estima que es 1 ($c'_i = 1$).

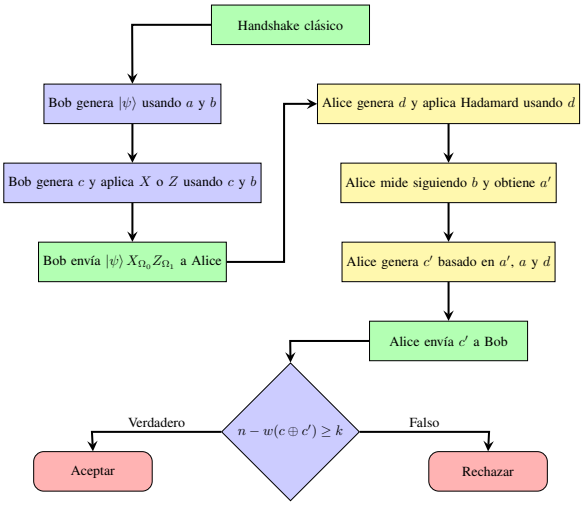


Figura 1. Diagrama de flujo de la QZKP propuesta

Ahora, se considera el caso en que $a'_i = a_i$. En este caso, Alice debe diferenciar si $i \in \Gamma$ o $i \notin \Gamma$. Si $i \notin \Gamma$ la única opción posible es que Bob no haya modificado el estado $|\psi\rangle$ y por tanto se hace la estimación $c'_i = 0$. En el caso de que $i \in \Gamma$ Alice es incapaz de inferir con seguridad si Bob a modificado el estado o no por el efecto de la codificación conjugada en consecuencia de la aplicación de la puerta H.

Así, resumiendo los resultados, Alice realiza la estimación de c'_i de la siguiente manera:

- Si $a'_i \neq a_i$, entonces $c'_i = d_i \oplus 1$.
- Si $a'_i = a_i$, entonces $c'_i = d_i$.

Nótese que $E[|\Gamma|] = \frac{n}{2}$ y que si $i \notin \Gamma$ Alice es capaz de acertar siempre el valor de c_i, sin embargo si $i \in \Gamma$ Alice acertará la estimación de c_i con probabilidad de $\frac{1}{2}$. Por tanto, Alice tiene un $75\,\%$ de probabilidad de acertar cada c_i.

Por otro lado, una probadora deshonesta que no conozca las bases de b sólo tiene una probabilidad de $50\,\%$ de adivinar correctamente cada bit de c. Por lo tanto, el porcentaje de acierto esperado de una probadora deshonesta sería de $50\,\%$.

La propiedad de conocimiento nulo está garantizada porque el verificador no tiene conocimiento del conjunto Γ elegido aleatoriamente por Alice. Si un verificador deshonesto intenta obtener información sobre las claves secretas a o b de Alice, podría preparar un estado cuántico específico para que la cadena devuelta c' proporcione información sobre las claves. Sin embargo, debido a la selección aleatoria de Γ, que permanece desconocida para el verificador, no se obtiene ninguna información.

Este protocolo no sigue estrictamente el paradigma clásico de una ZKP dado que se requiere la distribución previa de un secreto entre probadora y verificador, pero representa un avance hacia su consecución. Trabajos similares han intentado lograr el mismo objetivo utilizando diferentes enfoques [4].

En la práctica, los resultados no siempre se comportan como se espera, ya que las expectativas sólo representan promedios. Sin embargo, por la ley de los grandes números, este problema puede mitigarse aumentando n o repitiendo el protocolo varias veces. Probabilísticamente, aumentar n en un factor de k tiene el mismo efecto que repetir el protocolo k veces y calcular los aciertos medios de todas las ejecuciones. La seguridad de esta QZKP reside en gran medida en una buena elección del límite inferior para el porcentaje de coincidencia entre c y c', cercano a (pero inferior a) $75\,\%$.

III-D. Ejemplo Conceptual

Sean Alice y Bob dos participantes que comparten en secreto una secuencia de n bits y otra de n bases. Ambos utilizan dichas secuencias para generar una secuencia secreta compartida de n cúbits. La QZKP propuesta funciona de la siguiente manera:

1. Bob genera al azar una secuencia secreta de n bits, aplica por cada 1 de esa secuencia una puerta X o Z sobre el correspondiente cúbit de la secuencia compartida dependiendo de los valores de b, y envía a Alice los cúbits resultantes.
2. Alice genera al azar su propia secuencia secreta de n bits, aplica por cada 1 de esa secuencia una puerta Hadamard sobre el correspondiente cúbit recibido de Bob, y al final mide todos los cúbits utilizando la secuencia de n bases secretas compartidas. A partir de esa medición, Alice intenta deducir la secuencia secreta de Bob de la forma siguiente:
   - Lo único que sabe Alice con certeza es que, para cada cúbit en que ella no aplicó puerta Hadamard, si al medir obtuvo un bit distinto al bit secreto original, fue porque Bob sí aplicó una puerta X o Z sobre ese cúbit. Por tanto, en este caso, Alice puede concluir con certeza que ese bit de la secuencia secreta de Bob es 1.
   - Para cada cúbit en que Alice no aplicó puerta Hadamard, si al medir obtuvo el bit secreto original, Alice estima que ese bit de la secuencia secreta de Bob es 0.
   - Para cada cúbit en que Alice aplicó puerta Hadamard, si al medir obtuvo el bit secreto original, Alice estima que ese bit de la secuencia secreta de Bob es 1.
   - Para cada cúbit en que Alice aplicó puerta Hadamard, si al medir obtuvo un bit distinto al bit secreto original, Alice estima que ese bit de la secuencia secreta de Bob es 0.

A continuación se describe el análisis de los resultados:

- Alice no aplica puerta Hadamard a aproximadamente $n/2$ cúbits:
  - A aproximadamente la mitad de esos $n/2$ cúbits, Bob le aplicó puerta X o Z. Entre esos $n/4$ cúbits, se espera que el resultado de la medición sea distinto del bit original en la mitad de los casos. Por lo tanto, en aproximadamente $n/8$ bits, Alice acertará con certeza que el bit secreto de Bob es 1.
  - A los restantes (aproximadamente la mitad de los $n/2$ cúbits), Bob tampoco le aplicó ninguna puerta. Por tanto, en estos $n/4$ cúbits, el resultado de la medición de Alice será igual al bit original y Alice acertará que el bit secreto de Bob es 0.
- Alice aplica puerta Hadamard a los restantes (aproximadamente $n/2$ cúbits):

- A aproximadamente la mitad de esos $n/2$ cúbits, Bob también le aplicó una puerta X o Z. Entre esos $n/4$ cúbits, se espera que el resultado de la medición sea igual al bit original y Alice acertará que el bit secreto de Bob es 1.
- A los restantes (aproximadamente la mitad de los $n/2$ cúbits), Bob no aplicó ninguna puerta. Por tanto, en estos casos, Alice acertará por azar.

Por lo tanto, si Alice es honesta, podemos concluir que el porcentaje de aciertos de Alice sobre los bits secretos de Bob es aproximadamente el $75\,\%$.

IV. Seguridad del Protocolo

IV-A. Elección del Umbral

Teniendo en cuenta el ejemplo anterior, es posible determinar las probabilidades asociadas a las propiedades de completitud y solidez. Al establecer un umbral $k\,\%$, se puede calcular la probabilidad de que un probador deshonesto logre vulnerar el protocolo (solidez), así como la probabilidad de que un probador honesto sea incorrectamente rechazado (completitud). A continuación, en la Tabla I, se presentan estas probabilidades para distintos valores de umbral, considerando una configuración de 1024 cúbits.

Umbral (k %)	$P(X_{h\,\%} \leq k\,\%)$	$P(X_{d\,\%} \geq k\,\%)$
60	$3{,}284 \times 10^{-26}$	$9{,}809 \times 10^{-11}$
61	$2{,}918 \times 10^{-23}$	$1{,}315 \times 10^{-12}$
62	$1{,}728 \times 10^{-20}$	$1{,}178 \times 10^{-14}$
63	$1{,}207 \times 10^{-17}$	$4{,}113 \times 10^{-17}$
64	$3{,}015 \times 10^{-15}$	$1{,}556 \times 10^{-19}$
65	$4{,}964 \times 10^{-13}$	$3{,}872 \times 10^{-22}$
66	$5{,}367 \times 10^{-11}$	$6{,}302 \times 10^{-25}$
67	$5{,}674 \times 10^{-9}$	$3{,}281 \times 10^{-28}$
68	$2{,}503 \times 10^{-7}$	$2{,}144 \times 10^{-31}$
69	$7{,}162 \times 10^{-6}$	$8{,}974 \times 10^{-35}$
70	$1{,}325 \times 10^{-4}$	$2{,}388 \times 10^{-38}$

Tabla I
Probabilidades para solidez y completitud del protocolo según distintos umbrales con 1024 cúbits.

IV-B. Conocimiento Nulo

Si un verificador deshonesto intentara obtener la clave secreta del probador, podría preparar un estado constante $|\phi\rangle$ en cada intento de autenticación. De este modo, podría comparar de manera estratégica las estimaciones enviadas por la probadora con el objetivo de extraer información sobre la clave secreta. En este contexto, las modificaciones de conocimiento nulo aplicadas por el probador, denotadas como H_Γ, juegan un papel fundamental. Si el estado no ha sido preparado de acuerdo con las reglas del protocolo, dichas modificaciones, que varían aleatoriamente en cada intento de autenticación, introducen una aleatorización en la estimación c' de Alice, impidiendo así cualquier intento de inferencia por parte del verificador malicioso.

Es decir, si un verificador deshonesto intentara obtener información sobre la clave secreta del probador, podría preparar un estado de reto que no cumpla con las reglas del protocolo, en otras palabras, un estado trampa. Este estado se expresa como $|\phi\rangle = \alpha_0 |0\rangle + \alpha_1 |1\rangle$. Para analizar la posible filtración de información, se presentan los diagramas en árbol correspondientes a cada posible valor de la clave secreta del probador. A través de este análisis, se estudia la distribución de las respuestas posibles c'_i, lo que permite evaluar si existe una fuga de información. Cabe destacar que este estudio puede restringirse a la ejecución del protocolo para un solo cúbit, dado que cada cúbit se procesa de manera independiente de los demás. A continuación se discute cada caso posible.

El primer caso es cuando $b_i = C$ y $a_i = 0$. Al probador le llega el estado $|\phi\rangle$ preparado por el verificador (quien puede ser honesto o deshonesto). El probador, con probabilidad $\frac{1}{2}$, aplica una puerta Hadamard (H) o no aplica nada.

- Si no aplica la puerta Hadamard y mide en la base C:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 0$) con probabilidad $||\alpha_0||^2$.
  - Obtiene el resultado 1 (del cual se deduce $c'_i = 1$) con probabilidad $||\alpha_1||^2$.
- Si aplica la puerta Hadamard y mide en la base H:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 1$) con probabilidad $\left|\left|\frac{\alpha_0+\alpha_1}{\sqrt{2}}\right|\right|^2$.
  - Obtiene el resultado 1 (del cual se deduce $c'_i = 0$) con probabilidad $\left|\left|\frac{\alpha_0-\alpha_1}{\sqrt{2}}\right|\right|^2$.

Y por tanto:

- $P(c'_i = 0 \mid b_i = C \wedge a_i = 0) = \frac{1}{2}\left(||\alpha_0||^2 + ||\frac{\alpha_0-\alpha_1}{\sqrt{2}}||^2\right)$
- $P(c'_i = 1 \mid b_i = C \wedge a_i = 0) = \frac{1}{2}\left(||\alpha_1||^2 + ||\frac{\alpha_0+\alpha_1}{\sqrt{2}}||^2\right)$

El segundo caso es cuando $b_i = H$ y $a_i = 1$. Al probador le llega el estado $|\phi\rangle$ preparado por el verificador (quien puede ser honesto o deshonesto). El probador, con probabilidad $\frac{1}{2}$, aplica una puerta Hadamard (H) o no aplica nada.

- Si no aplica la puerta Hadamard y mide en la base H:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 1$) con probabilidad $\left|\left|\frac{\alpha_0+\alpha_1}{\sqrt{2}}\right|\right|^2$.
  - Obtiene el resultado 1 (del cual se deduce $c'_i = 0$) con probabilidad $\left|\left|\frac{\alpha_0-\alpha_1}{\sqrt{2}}\right|\right|^2$.
- Si aplica la puerta Hadamard y mide en la base H:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 0$) con probabilidad $||\alpha_0||^2$.
  - Obtiene el resultado 1 (del cual se deduce $c'_i = 1$) con probabilidad $||\alpha_1||^2$.

Y de aquí se deduce:

- $P(c'_i = 0 \mid b_i = H \wedge a_i = 1) = \frac{1}{2}\left(||\frac{\alpha_0-\alpha_1}{\sqrt{2}}||^2 + ||\alpha_0||^2\right)$
- $P(c'_i = 1 \mid b_i = H \wedge a_i = 1) = \frac{1}{2}\left(||\frac{\alpha_0+\alpha_1}{\sqrt{2}}||^2 + ||\alpha_1||^2\right)$

Entonces, de estos dos casos se obtiene:

$$P(c'_i = k \mid b_i = C \wedge a_i = 0) = P(c'_i = k \mid b_i = H \wedge a_i = 1) \tag{2}$$

siendo $k \in \{0, 1\}$.

El tercer caso posible es cuando $b_i = C$ y $a_i = 1$. Al probador le llega el estado $|\phi\rangle$ preparado por el verificador (quien puede ser honesto o deshonesto). El probador, con probabilidad $\frac{1}{2}$, aplica una puerta Hadamard (H) o no aplica nada.

- Si no aplica la puerta Hadamard y mide en la base C:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 1$) con probabilidad $||\alpha_0||^2$.

  - Obtiene el resultado 1 (del cual se deduce $c'_i = 0$) con probabilidad $||\alpha_1||^2$.
- Si aplica la puerta Hadamard y mide en la base H:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 0$) con probabilidad $\left|\left|\frac{\alpha_0+\alpha_1}{\sqrt{2}}\right|\right|^2$.
  - Obtiene el resultado 1 (del cual se deduce $c'_i = 1$) con probabilidad $\left|\left|\frac{\alpha_0-\alpha_1}{\sqrt{2}}\right|\right|^2$.

Siguiendo la misma lógica que anteriormente:

- $P(c'_i = 0 \mid b_i = C \wedge a_i = 1) = \frac{1}{2}\left(||\alpha_1||^2 + ||\frac{\alpha_0+\alpha_1}{\sqrt{2}}||^2\right)$
- $P(c'_i = 1 \mid b_i = C \wedge a_i = 1) = \frac{1}{2}\left(||\alpha_0||^2 + ||\frac{\alpha_0-\alpha_1}{\sqrt{2}}||^2\right)$

El último caso posible es cuando $b_i = H$ y $a_i = 0$. Al probador le llega el estado $|\phi\rangle$ preparado por el verificador (quien puede ser honesto o deshonesto). El probador, con probabilidad $\frac{1}{2}$, aplica una puerta Hadamard (H) o no aplica nada.

- Si no aplica la puerta Hadamard y mide en la base H:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 0$) con probabilidad $\left|\left|\frac{\alpha_0+\alpha_1}{\sqrt{2}}\right|\right|^2$.
  - Obtiene el resultado 1 (del cual se deduce $c'_i = 1$) con probabilidad $\left|\left|\frac{\alpha_0-\alpha_1}{\sqrt{2}}\right|\right|^2$.
- Si aplica la puerta Hadamard y mide en la base H:
  - Obtiene el resultado 0 (del cual se deduce $c'_i = 1$) con probabilidad $||\alpha_0||^2$.
  - Obtiene el resultado 1 (del cual se deduce $c'_i = 0$) con probabilidad $||\alpha_1||^2$.

Y de este último caso se deduce:

- $P(c'_i = 0 \mid b_i = H \wedge a_i = 0) = \frac{1}{2}\left(||\frac{\alpha_0+\alpha_1}{\sqrt{2}}||^2 + ||\alpha_1||^2\right)$
- $P(c'_i = 0 \mid b_i = H \wedge a_i = 0) = \frac{1}{2}\left(||\frac{\alpha_0-\alpha_1}{\sqrt{2}}||^2 + ||\alpha_0||^2\right)$

Entonces, de estos dos últimos casos se concluye:

$$P(c'_i = k \mid b_i = C \wedge a_i = 1) = P(c'_i = k \mid b_i = H \wedge a_i = 0) \tag{3}$$

donde $k \in \{0, 1\}$. De las ecuaciones 2 y 3 se demuestra que un verificador deshonesto es incapaz de obtener información sobre las claves privadas a y b. Aunque podría parecer que dicho verificador puede determinar si se encuentra en el caso 2 o 3, sigue estando limitado a encontrar una clave de n bits por fuerza bruta. Esto implica un coste computacional de 2^n, o bien $2^{\frac{n}{2}}$ si se considera el algoritmo de Grover [1].

Otra posible fuente de filtrado de información podría ser el *estado prueba* que envía Bob (en este caso verificador honesto) $|\psi\rangle X_{\Omega_0} Z_{\Omega_1}$, $\Omega_k = \{i \mid c_i = 1 \wedge b_i = k\}$. Es decir, si un probador deshonesto intentase autenticarse a Bob haciéndose pasar por Alice, podía obtener tantos *estados prueba* como quisiera, y por tanto medir el estado con respecto a bases de su elección. Sin embargo, las modificaciones que realiza Bob a cada cúbit tiene el efecto de aleatorizar el bit de salida, si se ha elegido correctamente la base, con probabilidad de $\frac{1}{2}$, y en cada iteración estas modificaciones se volvería a realizar o no aleatoriamente. Por tanto, ningún probador deshonesto podría ser capaz de obtener ninguna información de la clave secreta de Bob y Alice, tanto de la secuencia a como de la secuencia b.

V. Implementación del Protocolo

La implementación se ha llevado a cabo en el simulador cuántico Qiskit desarrollado por IBM [11]. La implementación se puede encontrar en el repositorio [12]. A continuación, en las figuras 2 y 3, se muestran los resultados de las implementaciones realizadas usando 512 cúbits y 1024 cúbits lógicos [13], es decir, no se produce absolutamente ningún error a lo largo de la ejecución del protocolo.

Se puede ver claramente la diferencia entre los intentos de probadores honestos contra los intentos de los probadores deshonestos. Además también es posible observar como la varianza de ambos casos disminuye a medida que se aumenta n, siguiendo las predicciones hechas en los apartados teóricos anteriormente expuestos.

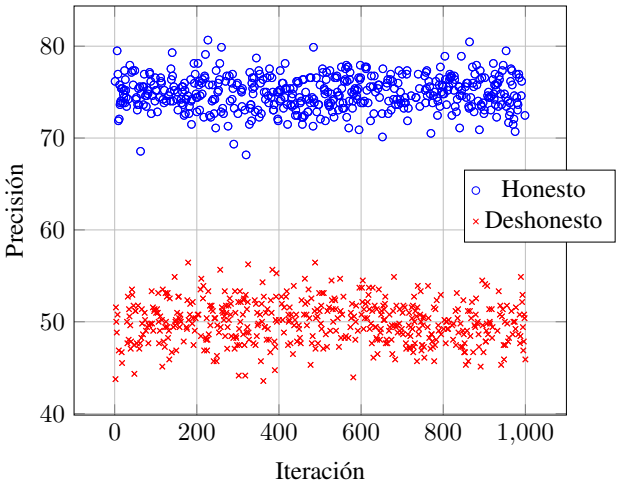


Figura 2. Resultados de la experimentación ideal (512 cúbits).

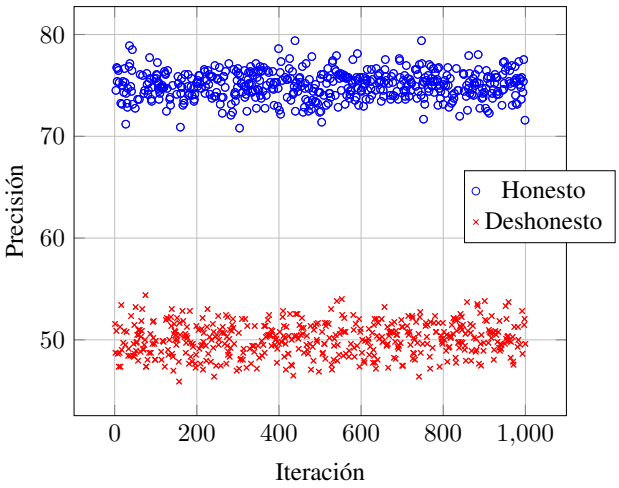


Figura 3. Resultados de la experimentación ideal (1024 cúbits).

VI. Conclusiones y Trabajos Futuros

En este trabajo se ha propuesto una nueva Prueba Cuántica de Conocimiento Nulo basada en codificación conjugada para la autenticación de usuarios sobre canales cuánticos. A diferencia de los esquemas clásicos de pruebas de conocimiento

nulo, que dependen de problemas matemáticos difíciles de resolver, este enfoque se basa en principios físicos fundamentales de la mecánica cuántica, garantizando la seguridad del protocolo sin asumir la dificultad computacional de ciertos problemas.

El análisis de la seguridad del protocolo ha permitido demostrar sus propiedades esenciales: completitud, solidez y conocimiento nulo. Se ha establecido que un usuario honesto tiene una probabilidad significativamente mayor de superar la prueba en comparación con un atacante que intente suplantar su identidad. Esto se refleja en los resultados de las distintas simulaciones realizadas, evidenciando que el protocolo puede diferenciar con alta confianza entre ambas partes.

Mediante simulaciones en Qiskit, se ha validado la viabilidad del protocolo. Los resultados muestran que el protocolo es claramente viable, manteniendo un buen rendimiento en la autenticación. Además, la flexibilidad del esquema permite ajustarlo a distintos niveles de recursos cuánticos disponibles, lo que facilita su implementación en hardware cuántico limitado.

Otro aspecto relevante de este trabajo es que el protocolo permite la autenticación sin que el verificador obtenga información adicional sobre la clave secreta del probador, garantizando la propiedad de conocimiento nulo. Esto impide que un verificador deshonesto obtenga ventajas indebidas o comprometa la seguridad del sistema.

Esta prueba de conocimiento nulo es una alternativa para la seguridad en redes cuánticas. Su solidez frente a ataques permite su integración con protocolos de Distribución Cuántica de Claves, mejorando la protección en entornos críticos como gobiernos, banca y defensa.

Como trabajo futuro, se propone su implementación en hardware cuántico real y el análisis de posibles ataques cuánticos basados en entrelazamiento contra su implementación, como por ejemplo la clonación imperfecta [14] y el posible uso de six-state para este protocolo. Otro ámbito de estudio posible es como combinar este protocolo en conjunto con protocolos de QKD. También se propone el estudio de nuevos protocolos cuánticos en los que no sea necesario que tanto probadora como verificador compartan una clave secreto compartida, sin basarse en problemas matemáticos computacionalmente difíciles.

Agradecimientos

Este trabajo ha sido posible gracias a los proyectos de investigación PID2022-138933OB-I00 y 2023DIG28 IACTA financiados por MCIN/AEI/ 10.13039/501100011033/FEDER EU, y la Fundación CajaCanarias la Caixa. También forma parte de las Cátedras de Ciberseguridad patrocinadas por Binter, y por INCIBE a través de una iniciativa llevada a cabo en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation).

Referencias

[1] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, ser. STOC '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 212–219. [Online]. Available: https://doi.org/10.1145/237814.237866

[2] P. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 1994, pp. 124–134.

[3] A. Broadbent, Z. Ji, F. Song, and J. Watrous, "Zero-knowledge proof systems for qma," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, Oct. 2016, p. 31–40. [Online]. Available: http://dx.doi.org/10.1109/FOCS.2016.13

[4] M. I. García Cid, D. Bodanapu, R. M. Sánchez-Ledesma, L. Ortiz Martín, V. Martín Ayuso, M. Brunero, A. Gatto, and P. Martelli, "Quantum zero-knowledge protocol for identity authentication," in *Quantum Engineering and Technology Conference (QET 2023)*, vol. 2023, 2023, pp. 15–18.

[5] S. Wiesner, "Conjugate coding," *SIGACT News*, vol. 15, no. 1, p. 78–88, Jan. 1983. [Online]. Available: https://doi.org/10.1145/1008908.1008920

[6] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theoretical Computer Science*, vol. 560, pp. 7–11, 2014, theoretical Aspects of Quantum Cryptography – celebrating 30 years of BB84. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304397514004241

[7] C. H. Bennett, G. Brassard, and N. D. Mermin, "Quantum cryptography without bell's theorem," *Phys. Rev. Lett.*, vol. 68, pp. 557–559, Feb 1992. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.68.557

[8] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof systems," *SIAM Journal on Computing*, vol. 18, no. 1, pp. 186–208, 1989. [Online]. Available: https://doi.org/10.1137/0218012

[9] P. A. M. Dirac, "A new notation for quantum mechanics," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 35, no. 3, p. 416–418, 1939.

[10] D. Escanez-Exposito, P. Caballero-Gil, M. Rodriguez-Vega, F. Costa-Cano, and E. Sáenz-de Cabezón, "Algebraic language for the efficient representation and optimization of quantum circuits," *Physica Scripta*, vol. 100, no. 2, p. 025107, jan 2025. [Online]. Available: https://dx.doi.org/10.1088/1402-4896/ad9fb6

[11] A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman, J. Gacon, S. Martiel, P. D. Nation, L. S. Bishop, A. W. Cross, B. R. Johnson, and J. M. Gambetta, "Quantum computing with Qiskit," 2024.

[12] J. Garcia-Diaz, "GitHub - jorgegardiaz/conjugate_coding_QZKP: Scripts con las simulaciones del protocolo de autenticación mediante una QZKP basada en codificación conjugada. — github.com," https://github.com/jorgegardiaz/conjugate_coding_QZKP, 2025.

[13] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*, 2010.

[14] V. Bužek and M. Hillery, "Quantum copying: Beyond the no-cloning theorem," *Physical Review A*, vol. 54, no. 3, p. 1844, 1996.

Towards IoT Security: A Hybrid Encryption Scheme based on Lightweight and Post-Quantum Cryptography

Lamine Syne
Universidad de La Laguna
Tenerife, España
lsyne@fg.ull.es

Édgar Pérez-Ramos
Universidad de La Laguna
Tenerife, España
eperezra@ull.edu.es

Candelaria Hernández-Goya
Universidad de La Laguna
Tenerife, España
mchgoya@ull.edu.es

Pino Caballero-Gil
Universidad de La Laguna
Tenerife, España
pcaballe@ull.edu.es

***Abstract*—The Internet of Things (IoT) connects over 30 billion devices worldwide, ranging from smartwatches to industrial equipment. However, securing communication in resource-constrained IoT devices remains a major challenge, as conventional cryptography becomes vulnerable in the quantum era. This paper proposes a hybrid encryption scheme that combines post-quantum key establishment (ML-KEM) and the lightweight authenticated cipher ASCON, both standardized by NIST (ASCON pending final approval). The proposed scheme addresses the dual challenges of post-quantum security and constrained resources. Experimental evaluations demonstrate its feasibility and efficiency in typical IoT environments, supporting robust and scalable secure communication.**

***Index Terms*—IoT Security, Lightweight Cryptography, Post-quantum protocols**

I. Introduction

When deploying IoT solutions, it is essential to consider several critical factors inherent to the devices, such as processing power, memory constraints, energy efficiency, connectivity, data processing capabilities, and security and privacy requirements. Properly addressing these aspects is crucial to foster widespread adoption and trust across domains such as transportation, smart homes, healthcare, and industrial systems. To achieve both security and efficiency in IoT environments, lightweight cryptographic algorithms play a pivotal role in accommodating the resource limitations of these devices.

Conventional cryptographic techniques, based on mathematically hard problems, are increasingly considered vulnerable in light of quantum computing advancements [1]. Early recommendations, such as those outlined in [2], proposed increasing key sizes in symmetric cryptography to mitigate quantum threats. However, the adoption of quantum-resistant solutions in resource-constrained IoT environments introduces additional challenges. For instance, Quantum Key Distribution (QKD) has been proposed as a potential alternative, but its reliance on complex and costly infrastructure, combined with limited range and scalability, makes it currently impractical for widespread IoT deployments [3].

The latest Post-Quantum Cryptographic (PQC) algorithms standardized by NIST (see Section II-B) include Public Key Cryptography (PKC) solutions designed for secure key distribution across various platforms. This study focuses on analyzing the applicability of these solutions in IoT environments, specifically addressing the constraints posed by resource-limited devices. We evaluate the Key Encapsulation Mechanism (KEM) defined in [4], leveraging its implementation through the Open Quantum Safe (OQS) project [5], to establish quantum-resistant key distribution for IoT devices.

In addition, we employ lightweight authenticated encryption from the ASCON family to secure device-to-device communication channels, providing confidentiality, integrity, and authenticity. Although ASCON-80pq is not officially included among the finalized variants in the latest NIST draft, we incorporate it into our evaluation as a candidate optimized for post-quantum and resource-constrained settings, and compare its performance and security properties with ASCON-AEAD128, the NIST-recommended variant.

Our findings confirm the practicality of deploying quantum-resistant cryptographic solutions within real-world IoT scenarios, demonstrating a viable pathway for achieving secure and efficient communication in the post-quantum era.

The remainder of this paper is organized as follows. The remainder of this paper is organized as follows: Section II introduces foundational concepts, including IoT security issues, Lightweight and post-quantum and NIST-standardized cryptographic solutions ML-KEM and ASCON. Section III reviews recent advancements in PQC for IoT, focusing on practical implementation challenges. The proposed architecture, which integrates quantum-resistant cryptographic mechanisms, is described in Section IV. Section V and presents the experimental setup. Finally, Section VI presents performance evaluation of the proposed scheme, analyzing key security and efficiency parameters, summarizes the main findings, evaluates the feasibility of the proposed solution, and outlines future directions for advancing quantum-resistant IoT security.

II. Background

A. Lightweight Cryptography

An IoT ecosystem relies on a wide variety of communication protocols, each with unique characteristics, standards, and vulnerabilities that require tailored security measures. This diversity, combined with the need for interoperability, significantly increases the attack surface. Furthermore, resource constraints in many IoT devices, such as limited processing power, memory, and energy, hinder the deployment of conventional robust security mechanisms, complicating system protection and risk management. Therefore, adopting

lightweight cryptography (LWC) is essential to address these challenges.

In February 2023, NIST selected the *ASCON* family of algorithms for standardization in lightweight cryptographic applications [6]. Later, in November 2024, NIST released the initial public draft of Special Publication 800-232 [7], which introduces a family of symmetric key cryptographic primitives based on ASCON. Open for public comment until February 2025, this draft defines *ASCON-AEAD128*, *ASCON-Hash256*, *ASCON-XOF128*, and *ASCON-CXOF128*, providing functionalities such as Authenticated Encryption with Associated Data (AEAD), a hash function, and two Extendable Output Functions (XOFs). The ASCON family is optimized for efficiency and security in resource-constrained environments.

Although *ASCON-80pq* is not included in the final NIST-standardized ASCON variants, it was originally proposed as a lightweight authenticated encryption scheme with post-quantum considerations. ASCON-80pq uses a 160-bit key, a 128-bit nonce, and a 128-bit authentication tag. It provides approximately 80-bits of post-quantum security. Its design aims to balance resource efficiency and a basic level of quantum-resistant security, making it suitable for severely constrained IoT devices where implementing full 128-bit post-quantum cryptography may be impractical. In this work, we include *ASCON-80pq* in our evaluation to explore its potential as a lightweight alternative and to analyze security-performance trade-offs when compared to the officially standardized *ASCON-AEAD128*, providing valuable insights for post-quantum IoT scenarios.

B. Post-Quantum Cryptography

The National Institute of Standards and Technology (NIST) has recently developed the corresponding draft of the Federal Information Processing Standards (FIPS), *FIPS 203* [4], specifying the Module-Lattice Key-Encapsulation Mechanism (ML-KEM) with Indistinguishability under Chosen Ciphertext Attack (IND-CCA2) security.

In a typical KEM protocol, two parties, Alice and Bob, securely establish a shared secret key.

According to [8], a Key Encapsulation Mechanism (KEM) is a tuple of probabilistic algorithms $(KGen, Enc, Dec)$:

- $KGen$: key generation algorithm producing a key pair (sk, pk), where sk is a secret key and pk is a public key.
- Enc: encapsulation algorithm that takes a public key pk and produces a pair (c, K), where K is a shared secret and c is the public encapsulation.
- Dec: decapsulation algorithm that takes a secret key sk and a ciphertext c to recover the shared secret K.

Alice runs the key generation algorithm $KGen()$ to produce a public and secret key pair (pk, sk). Bob then executes the encapsulation algorithm $Enc(pk)$, generating a ciphertext c and a shared secret K, i.e., $Enc(pk) = (c, K)$. Upon receiving c, Alice runs the decapsulation algorithm $Dec(sk, c)$ to recover K. At the conclusion of this protocol, both parties share the same secret K.

The parameters used in ML-KEM are shown in Table I, and the corresponding key and ciphertext sizes are detailed in Table II. Furthermore, as shown in [8], the algorithms Enc and Dec may fail to produce a valid output, denoted by the symbol $\perp$. In the case of ML-KEM, these failures are extremely rare because the parameters are carefully designed to minimize the probability of decapsulation failure (see Table III, where the failure rate is lower than 2^{-100} in all cases).

Table I
PARAMETER SETS FOR ML-KEM

Algorithms	n	k	q	η_1	η_2	(d_u, d_v)
ML-KEM-512	256	2	3329	3	2	(10, 4)
ML-KEM-768	256	3	3329	2	2	(10, 4)
ML-KEM-1024	256	4	3329	2	2	(11, 5)

Table II
SIZES (IN BYTES) OF KEYS AND CIPHERTEXTS OF ML-KEM

Algorithms	Encap. key	Decap. key	Ciphertext
ML-KEM-512	800	1632	768
ML-KEM-768	1184	2400	1088
ML-KEM-1024	1568	3168	1568

Table III
DECAPSULATION FAILURE RATES FOR ML-KEM

Parameter set	Decapsulation failure rate
ML-KEM-512	$2^{-138.8}$
ML-KEM-768	$2^{-164.8}$
ML-KEM-1024	$2^{-174.8}$

ML-KEM is derived from CRYSTALS-Kyber, incorporating several optimizations to enhance efficiency and security. Like Kyber, ML-KEM employs the Fujisaki-Okamoto (FO) transformation to ensure resistance against chosen-ciphertext attacks (CCA). However, ML-KEM modifies the application of FO by removing the ciphertext hashing step in shared secret derivation, eliminating redundant randomness hashing, and introducing explicit input validation. These modifications reduce computational overhead while maintaining strong security guarantees, making ML-KEM more efficient and robust for resource-constrained environments such as IoT.

The *Open Quantum Safe* (OQS) project [9] and its *liboqs* library [10] offer implementations of quantum-resistant cryptographic algorithms, including ML-KEM, with standard APIs, test harnesses, and benchmarking tools. OQS also provides integration with libraries like OpenSSL, enabling practical evaluation of post-quantum algorithms.

Together, ML-KEM and ASCON address key IoT security needs: quantum-resistant key exchange and lightweight authenticated encryption, offering secure and efficient solutions tailored to the constraints of IoT environments.

III. RELATED WORK

Aware of the growing need to develop cryptographic mechanisms that resist quantum threats, particularly for resource-constrained IoT devices, recent efforts have focused on evaluating and optimizing Post-Quantum Cryptographic solutions.

An updated survey on this topic is available in [11], which reviews 35 recent studies on PQC for IoT, focusing on performance evaluation, software and hardware optimizations, and GPU acceleration. Most works emphasize NIST-approved algorithms, such as Kyber and Dilithium, and highlight the challenges of balancing speed, memory usage, and energy efficiency in constrained environments. Hardware-based approaches using FPGAs (Field Programmable Gate Arrays) and ASICs (Application Specific Integrated Circuits) are included, while software efforts focus on parallelization and optimized polynomial arithmetic. Nevertheless, the absence of standardized evaluation frameworks and incomplete alignment with NIST standards remain obstacles for widespread IoT adoption.

The study in [12] proposes a framework to select tailored security solutions for post-quantum threats, enabling researchers and practitioners to choose the most appropriate cryptographic mechanisms based on IoT-specific requirements and attack scenarios.

Several studies have analyzed the impact of PQC on IoT devices. For example, [13] examines the feasibility of integrating post-quantum algorithms, including NTRU, into platforms equipped with low-power microcontrollers, such as ESP32 and Raspberry Pi. The work also evaluates combinations of key encapsulation mechanisms (CRYSTALS-Kyber, NTRU, Saber) and digital signatures (CRYSTALS-Dilithium, Falcon, SPHINCS+) to assess their impact on dense IoT networks. Similarly, [14] compares three KEM schemes based on runtime and memory consumption, concluding that Kyber512 is the most energy-efficient solution for devices such as wearables, environmental sensors, and smart home devices. In line with these findings, [15] presents experiments on fog nodes and IoT devices (Raspberry Pi and ESP32 running real-time operating system), showing that although Kyber-512 achieves the fastest execution, Kyber-768 offers the best trade-off between speed, stability, and security, making it the most suitable option for general IoT applications. Conversely, while Kyber-1024 provides the highest security, it suffers from slower performance and greater instability, making it unsuitable for time-critical applications.

The combination of KEM and lightweight cryptography (LWC) for IoT has been previously explored. For instance, [16] presents a KEM scheme optimized for hardware, specifically targeting FPGA-based platforms to achieve high performance in constrained environments. Similarly, [17] integrates CRYSTALS-Kyber and ASCON in an Internet of Medical Things (IoMT) system managed via Kubernetes (K3S), demonstrating a practical post-quantum security solution for healthcare-focused IoT deployments.

While these works highlight the potential of combining KEM and LWC in specialized applications, they primarily focus on hardware-optimized implementations or specific domains such as IoMT. In contrast, our approach provides a broader, flexible, and scalable post-quantum security architecture, leveraging ML-KEM and ASCON in a generalized IoT setting.

Furthermore, combining hardware and software optimizations has proven effective in enhancing code-based cryptography. A co-design approach for the Hamming Quasi-Cyclic (HQC) KEM, tailored for IoT applications, is presented in [18]. By integrating software optimizations with hardware acceleration, this approach significantly enhances energy efficiency and computational performance.

Lastly, [19] evaluates the integration of PQC algorithms into TLS handshakes over Bluetooth Low Energy (BLE) and Wi-Fi. The study identifies Falcon-512 and Kyber512 as the most suitable digital signature and KEM schemes, respectively, for secure IoT communications. Although BLE connections exhibited higher delays due to bandwidth limitations, Wi-Fi networks efficiently handled larger signature schemes such as Dilithium2. The authors also propose a dynamic KEM selection mechanism that adapts to the communication protocol, further supporting the viability of Falcon-512 and Kyber512 for a quantum-resistant IoT ecosystem.

A. *Contributions*

Despite extensive research on PQC for IoT, as summarized in the previous section, several challenges remain unaddressed.

To overcome these limitations, this paper introduces a novel lightweight cryptographic scheme that integrates the standardized ML-KEM for post-quantum key establishment with the ASCON cipher family for authenticated encryption. The key contributions of this work include:

- Designing and evaluating a cryptographic scheme that combines ML-KEM and ASCON to secure IoT communications under resource constraints, enabling post-quantum secure key exchange and authenticated encryption for constrained devices.
- Comparing ASCON-AEAD128 and ASCON-80pq to examine the trade-offs between security and efficiency in post-quantum IoT environments, offering new insights into the applicability of ASCON-80pq for highly constrained devices.
- Conducting practical experiments using standardized implementations (via the OQS/liboqs project and official ASCON libraries) to assess key performance metrics, including execution time, memory consumption, and communication overhead, on resource-limited platforms.
- Evaluating the scheme's suitability for unidirectional, low-interaction key exchange and secure communication, which are critical in typical IoT use cases where device interaction is minimal and energy efficiency is paramount.

These contributions address key open questions in the current literature by providing a complete, standard-aligned, and experimentally validated approach to post-quantum secure communication for IoT ecosystems.

IV. Lightweight Quantum-Resistant Cryptographic Framework for IoT

Traditional security architectures for IoT systems typically integrate a key agreement mechanism, a Key Derivation Function (KDF), and authenticated encryption to ensure confidentiality, integrity, and authenticity of communications. In this work, we propose a lightweight, quantum-resistant security architecture specifically designed for resource-constrained IoT environments. Our framework combines ML-KEM for efficient post-quantum key agreement, an ASCON-based KDF,

and the ASCON family of authenticated ciphers for securing data transmissions. The overall design is illustrated in Fig. 1.

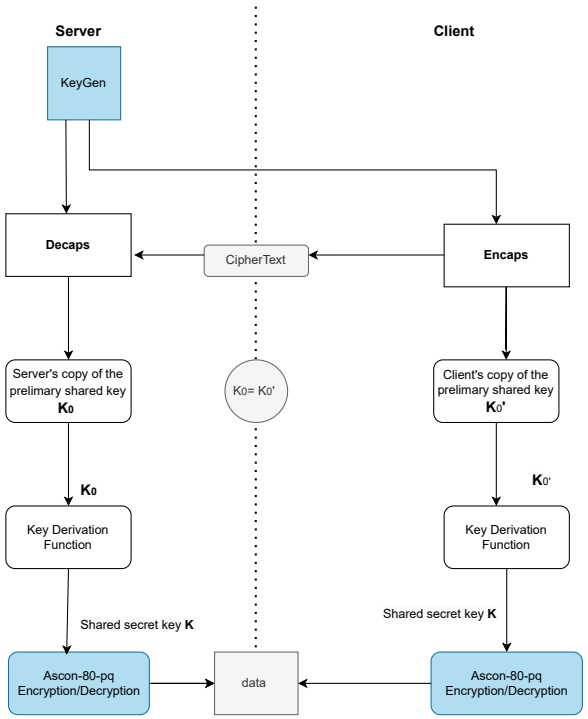


Figure 1. Proposed post-quantum security architecture combining ML-KEM and ASCON.

The proposed architecture leverages ML-KEM to establish a shared secret between IoT devices directly within the communication flow, minimizing computational overhead and simplifying key management. The established shared secret is then processed through an ASCON-based KDF, optimized for constrained devices, to derive cryptographic keys used for both encryption and authentication.

After the key encapsulation process, the shared secret obtained from the KEM undergoes further processing to enhance security and adaptability within cryptographic protocols. While the raw KEM output provides a strong initial key, it may not exhibit the necessary properties for direct use in encryption. To address this, we employ an ASCON-based Key Derivation Function (KDF), ensuring that the derived keys achieve high entropy, uniform randomness, and resistance against cryptographic attacks.

The integration of this KDF offers multiple security benefits. First, it removes any inherent structure in the KEM output that could be exploited through cryptanalysis. Additionally, it mitigates potential vulnerabilities to side-channel attacks by ensuring that the derived keys exhibit high entropy and uniform randomness. Furthermore, the ASCON-based KDF enhances forward secrecy by enabling periodic key refreshing through a rolling key mechanism, where each new session key is derived from the previous one, or by expanding the initial KEM-derived secret into multiple session keys using the ASCON-XOF (Extendable Output Function), ensuring continuous cryptographic security and efficient key diversification.

Data confidentiality, integrity, and authenticity are provided by the ASCON family of authenticated ciphers, which implement Authenticated Encryption with Associated Data (AEAD), protecting both payload and associated metadata. Due to ASCON's high efficiency and strong resistance to side-channel attacks, it is particularly well-suited for low-power IoT environments.

By integrating quantum-resistant primitives in a modular and scalable architecture, our framework delivers end-to-end security while ensuring interoperability and low resource consumption. This design offers a future-proof solution, addressing emerging post-quantum security needs in diverse and heterogeneous IoT ecosystems.

V. Experimental Setup

This section describes the experimental framework designed to evaluate the feasibility and performance of integrating post-quantum and lightweight cryptographic solutions in IoT environments. The experiments focus on measuring the runtime performance of key cryptographic mechanisms under various configurations. The hardware and software setup is optimized for resource-constrained environments.

Performance metrics, including encryption and decryption time, computational efficiency, and the impact of ciphertext and key sizes, are analyzed to assess the practicality of these solutions for real-world IoT scenarios. The study also evaluates the trade-offs between security and performance, highlighting the balance between execution time and cryptographic strength across different message sizes and key configurations.

A. Hardware and Software Configuration

The hardware platform selected for the experiments is the Raspberry Pi 4 Model B, with the following specifications:

- **CPU:** 64-bit quad-core Cortex-A72.
- **Memory:** 8 GB LPDDR4 RAM.
- **CPU Max Frequency:** 1.5 GHz.
- **Connectivity:** 2.4 GHz and 5.0 GHz IEEE 802.11b/g/n/ac wireless LAN, Bluetooth 5.0, BLE.

The software stack includes the following tools and libraries:

- **liboqs** [10]: A C library implementing quantum-resistant cryptographic algorithms.
- **liboqs-python** [20]: A Python 3 wrapper for liboqs, tested on Linux, macOS, and Windows platforms.
- **Python sockets**: Used to establish secure communication channels between devices.
- **Wireshark**: Captures and analyzes network traffic to verify correct data transmission, ensuring that exchanged messages and shared keys are encrypted and not exposed in plaintext.
- **PySASCON**: A Python library implementing the ASCON cipher suites.

B. Network Architecture

The experimental network is based on a device-to-device communication model composed of two Jetson Orin Nano modules. As illustrated in Fig. 2, the devices communicate over a Wi-Fi network and are powered through dedicated power cables. An access point (AP) router provides connectivity between the two nodes.

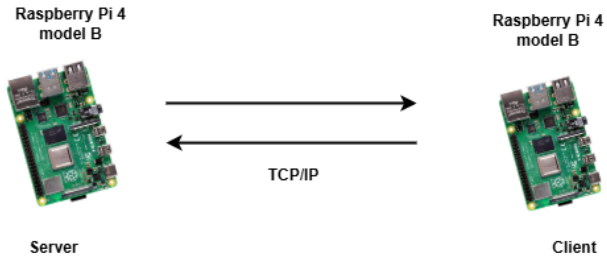


Figure 2. Experimental setup and network architecture

VI. Performance Analysis

1) Key Encapsulation Mechanism (KEM): The performance of the NIST-standardized ML-KEM algorithms (ML-KEM-512, ML-KEM-768, and ML-KEM-1024) was evaluated on the testbed. Table IV and Fig. 3 summarize the runtime for key generation, encapsulation, and decapsulation operations.

Table IV
KEM RUNTIME (MS)

Algorithm	Key Gen. (ms)	Encap. (ms)	Decap. (ms)
ML-KEM-512	0.2490	0.4535	0.3481
ML-KEM-768	0.4073	0.5127	0.4004
ML-KEM-1024	0.4658	0.5130	0.5931

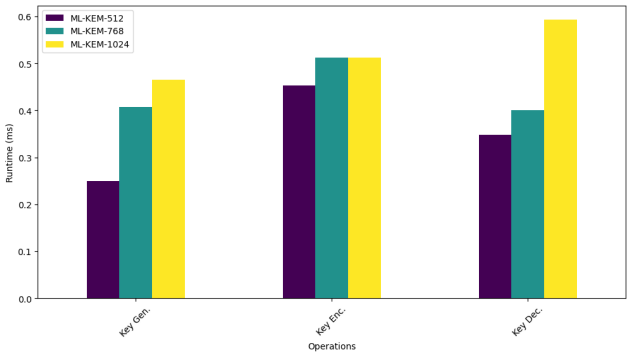


Figure 3. Performance of ML-KEM schemes.

Results show that ML-KEM-512 achieves the fastest performance across all operations, making it attractive for highly constrained devices, though it offers lower security. ML-KEM-768 emerges as the best compromise between performance and security, showing moderate runtimes and balanced resource demands. Conversely, while ML-KEM-1024 provides the highest security, its larger keys and ciphertext sizes result in higher latency, particularly noticeable in decapsulation times.

A. Encryption Schemes: ASCON-80pq and ASCON-AEAD128

To assess ASCON's performance, experiments were conducted to measure encryption and decryption runtimes for message sizes of 10 bytes, 100 bytes, and 500 bytes across various ML-KEM variants.

Table V
ASCON-80PQ ENCRYPTION AND DECRYPTION TIME (MS)

K. Establishment	M.length	Encrypt (ms)	Decrypt (ms)
ML-KEM-512	10	3.7208	2.5308
	100	4.2424	3.3999
	500	9.4190	7.2171
ML-KEM-768	10	2.5395	2.5951
	100	3.3916	3.5304
	500	7.7566	8.0373
ML-KEM-1024	10	2.9504	3.4750
	100	3.8088	4.6912
	500	9.8833	10.4243

Table VI
ASCON-AEAD128 ENCRYPTION AND DECRYPTION TIME (MS)

K. Establishment	M. length (B)	Encrypt (ms)	Decrypt (ms)
ML-KEM-512	10	3.1075	3.0606
	100	3.6724	3.6698
	500	8.0200	8.7566
ML-KEM-768	10	2.9550	2.8198
	100	3.6760	3.6284
	500	7.9112	7.9714
ML-KEM-1024	10	2.7311	3.4535
	100	3.5588	4.7131
	500	7.7034	8.6906

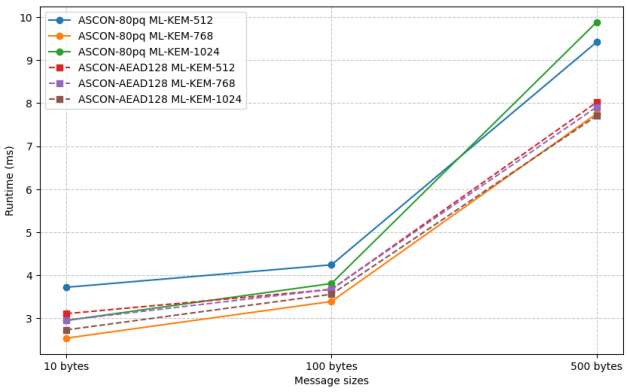


Figure 4. ASCON-80pq vs ASCON-AEAD128 encryption runtime

According to Table V, the encryption time for ASCON-80pq increases proportionally with message size, as larger messages require more computational resources. Among the ML-KEM variants, ML-KEM-512 consistently exhibits the longest encryption times, despite having a smaller key size. This behavior may be attributed to computational overhead introduced by authenticated encryption (AE) mechanisms, such as authentication tag generation and integrity verification. These additional operations contribute to the overall runtime and may explain why ML-KEM-512 does not necessarily outperform larger key variants in encryption speed.

ML-KEM-768 achieves the best balance between security and performance, offering the fastest encryption times across all tested message sizes. In contrast, ML-KEM-1024 incurs slightly higher computational overhead, reflecting the expected trade-off between cryptographic strength and execution

speed.

Similarly, decryption times follow a comparable trend, with ML-KEM-768 demonstrating the most efficient performance, while ML-KEM-512 and ML-KEM-1024 show relatively higher decryption latency. The additional processing required for authentication and integrity verification in AE schemes contributes to this variation. This highlights the impact of key size and authenticated encryption operations on cryptographic performance and reinforces the importance of selecting appropriate parameter sets for constrained IoT environments.

A comparative analysis of ASCON-AEAD128 and ASCON-80pq (see Figure 4) reveals significant differences in encryption and decryption times depending on message size and the ML-KEM variant used. Table VI presents the measured encryption and decryption times for ASCON-AEAD128, showing a similar trend to ASCON-80pq. For small message sizes (10 B and 100 B), ASCON-80pq achieves faster decryption times, particularly when combined with ML-KEM-512 and ML-KEM-768. However, as the message size increases (500 B), ASCON-AEAD128 demonstrates greater efficiency, offering more stable execution times, particularly when paired with ML-KEM-1024.

In terms of encryption performance, Table VI shows that ASCON-AEAD128 is slightly more efficient than ASCON-80pq in most cases, except for ML-KEM-768, where their performance remains comparable.

In conclusion, ASCON-80pq may be preferable for applications requiring rapid decryption and handling of small data volumes, whereas ASCON-AEAD128 proves to be more efficient for larger message sizes, ensuring a more balanced performance between encryption and decryption. The combination of ML-KEM-768 and ASCON-80pq presents the best trade-off between computational efficiency and post-quantum security, making it a strong candidate for constrained IoT environments.

B. Rationale for Evaluating ASCON-80pq

Although ASCON-80pq was initially proposed as a post-quantum variant of the ASCON family, it was not included in the final standardization by NIST. Although ASCON-80pq uses a 160-bit key, its effective post-quantum security level was estimated to be approximately 80-bits due to Grover's algorithm. This security margin was considered insufficient for standardization, motivating the limitation for variants like ASCON-AEAD128 in the final NIST recommendations.

Despite this, we include ASCON-80pq in our evaluation for the following reasons:

- It provides an interesting reference point for analyzing trade-offs between security and efficiency in resource-constrained environments.
- It allows us to study the impact of using a reduced key size in lightweight post-quantum encryption and compare its performance to ASCON-AEAD128.
- While no longer considered for standardization, ASCON-80pq represents a historical design choice aimed at balancing efficiency and minimal post-quantum security, offering valuable insights into the evolution of lightweight cryptographic standards.

By comparing ASCON-80pq with ASCON-AEAD128, we evaluate the practical impact of these design choices in real-world IoT scenarios. Our results examine whether transitioning to 128-bit keys significantly affects performance and whether ASCON-80pq retains any advantages for resource-constrained devices despite its reduced security margin. This analysis provides deeper insights into the trade-offs between lightweight cryptographic efficiency and post-quantum security. requirements.

CONCLUSIONS

This paper presents a hybrid and lightweight encryption scheme for securing IoT communications. The study demonstrates the feasibility of integrating Post-Quantum Cryptographic mechanisms into resource-constrained IoT systems. By leveraging ML-KEM for quantum-resistant key exchange and ASCON for lightweight authenticated encryption, the proposed architecture effectively addresses the challenges posed by quantum computing while maintaining compatibility with the computational and energy limitations of IoT devices.

Experimental results confirm that ML-KEM-768 with ASCON-80pq offers an effective balance between security and performance, making it well-suited for practical IoT applications. Furthermore, ASCON-AEAD128 ensures minimal computational overhead, enhancing data protection without compromising device responsiveness.

Future work will focus on optimizing the architecture to support a broader range of cryptographic primitives and further improving energy efficiency. Additionally, real-world IoT testing, incorporating heterogeneous devices and dynamic network conditions, will be crucial for validating scalability and adaptability. Advancing these directions will contribute to the development of quantum-resistant IoT systems capable of addressing emerging security challenges in the post-quantum era.

ACKNOWLEGMENTS

This work was partially funded by the Project SCITALA C064/23 ULL-INCIBE, the Cybersecurity Chairs of the University of La Laguna of INCIBE and Binter and EU Next Generation through Recovery, Transformation and Resilience Plan and the grant PID2022-138933OB-I00 funded by MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR.

REFERENCES

[1] A. Lohachab, A. Lohachab, and A. Jangra, "A comprehensive survey of prominent cryptographic aspects for securing communication in post-quantum iot networks," *Internet of Things*, vol. 9, p. 100174, 2020, https://doi.org/10.1016/j.iot.2020.100174.

[2] L. Chen, S. Jordan, Y.-K. Liu, D. Moody, R. Peralta, R. Perlner, and D. Smith-Tone, "Report on post-quantum cryptography," National Institute of Standards and Technology (NIST), Tech. Rep. NISTIR 8105, 2016, available free of charge from: http://dx.doi.org/10.6028/NIST.IR.8105. [Online]. Available: http://dx.doi.org/10.6028/NIST.IR.8105

[3] S. P C, K. Jain, and P. Krishnan, "Analysis of post-quantum cryptography for internet of things," in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2022, pp. 387–394, https://doi.org/10.1109/ICICCS53718.2022.9787987.

[4] National Institute of Standards and Technology, "Module-lattice-based key encapsulation mechanism standard," Department of Commerce, Washington, D.C., Tech. Rep. Federal Information Processing Standards Publication (FIPS) NIST FIPS 203 ipd, 2023, https://doi.org/10.6028/NIST.FIPS.203.ipd.

[5] D. Stebila and M. Mosca, "Post-quantum key exchange for the internet and the open quantum safe project," in *Selected Areas in Cryptography (SAC) 2016*, ser. Lecture Notes in Computer Science, R. Avanzi and H. Heys, Eds., vol. 10532. Springer, October 2017, pp. 1–24, . [Online]. Available: https://openquantumsafe.org

[6] "NIST Lightweight Cryptography competition (2019–2023)," 2024, . [Online]. Available: https://www.nist.gov/news-events/news/2023/02/nist-selects-lightweight-cryptography-algorithms-protect-small-devices

[7] M. S. Turan, K. McKay, D. Chang, J. Kang, and J. Kelsey, "Ascon-based lightweight cryptography standards for constrained devices: Authenticated encryption, hash, and extendable output functions," National Institute of Standards and Technology (NIST), NIST Special Publication (SP) 800-232 (Initial Public Draft), November 2024, comments Due: February 7, 2025. Email Comments to: sp800-232-comments@list.nist.gov. [Online]. Available: https://www.nist.gov/document/sp800-232-initial-public-draft

[8] M. Campagna and A. Petcher, "Security of hybrid key encapsulation," Cryptology ePrint Archive, Paper 2020/1364, 2020, https://eprint.iacr.org/2020/1364. [Online]. Available: https://eprint.iacr.org/2020/1364

[9] D. Stebila and M. Mosca, "Post-quantum key exchange for the internet and the open quantum safe project," in *Selected Areas in Cryptography – SAC 2016*, ser. Lecture Notes in Computer Science, vol. 10532. Springer, 2016, pp. 1–24, . [Online]. Available: https://eprint.iacr.org/2016/1017

[10] Open Quantum Safe Project, "liboqs: Open quantum safe c library," https://github.com/open-quantum-safe/liboqs, 2024, accessed: December 18, 2024.

[11] T. Liu, G. Ramachandran, and R. Jurdak, "Post-quantum cryptography for internet of things: A survey on performance and optimization," 2024, . [Online]. Available: https://arxiv.org/abs/2401.17538

[12] A. Alomari and S. A. Kumar, "Securing iot systems in a post-quantum environment: Vulnerabilities, attacks, and possible solutions," *Internet of Things*, vol. 25, pp. 101–132, 2024, https://doi.org/10.1016/j.iot.2024.101132.

[13] J. Señor Sánchez, "Analysis and evaluation of the impact of post-quantum cryptography at the edge of iot," 2024, no Publicado. [Online]. Available: https://oa.upm.es/81443/

[14] A. K. Kwala, S. Kant, and A. Mishra, "Comparative analysis of lattice-based cryptographic schemes for secure iot communications," *Discover Internet of Things*, vol. 4, no. 1, p. 13, Sep 2024, . [Online]. Available: https://doi.org/10.1007/s43926-024-00069-2

[15] R. Ristov and S. Koceski, "Quantum resilient public key cryptography in internet of things," in *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, 2023, pp. 1–4, https://doi.org/10.1109/MECO58584.2023.10154994.

[16] S. Kundu, A. Ghosh, A. Karmakar, S. Sen, and I. Verbauwhede, "Rudraksh: A compact and lightweight post-quantum key-encapsulation mechanism," *Cryptology ePrint Archive*, 2024, https://doi.org/10.48550/arXiv.2501.13799.

[17] E.-H. Mohamed, P. V. Ankunda, J. Ung, and W.-M. Hwu, "Securing the internet of medical things (iomt) with k3s and hybrid cryptography: Integrating post-quantum approaches for enhanced embedded system security," in *2024 IEEE 17th Dallas Circuits and Systems Conference (DCAS)*. IEEE, 2024, pp. 1–6, https://doi.org/10.1109/DCAS61159.2024.10539887.

[18] M. Schöffel, J. Feldmann, and N. Wehn, "Code-based cryptography in iot: A hw/sw co-design of hqc," 2023, . [Online]. Available: https://arxiv.org/abs/2301.04888

[19] J. Bozhko, Y. Hanna, R. Harrilal-Parchment, S. Tonyali, and K. Akkaya, "Performance evaluation of quantum-resistant tls for consumer iot devices," in *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, 2023, pp. 230–235, https://doi.org/10.1109/CCNC51644.2023.10060762.

[20] Open Quantum Safe Project, "liboqs-python: Python 3 bindings for liboqs," https://github.com/open-quantum-safe/liboqs-python, 2024, accessed: December 18, 2024.

Protección de Inferencias en Redes Neuronales Profundas usando Cifrado Homomórfico

Marcos Rodriguez-Vega
Universidad de La Laguna
San Cristóbal de La Laguna
mrodrive@ull.edu.es

Pino Caballero-Gil
Universidad de La Laguna
San Cristóbal de La Laguna
pcaballe@ull.edu.es

***Resumen*—El uso generalizado de redes neuronales profundas en entornos sensibles, como el sector sanitario o financiero, ha impulsado el desarrollo de métodos criptográficos para garantizar la privacidad, destacando especialmente el cifrado homomórfico, que permite realizar algunas operaciones sobre datos cifrados sin necesidad de descifrarlos. Este trabajo analiza la viabilidad y el rendimiento de las inferencias en redes neuronales que utilizan funciones de activación aproximadas mediante polinomios con el fin de aprovechar las propiedades de los cifrados homomórficos. Concretamente se estudia aquí el caso de uso de los polinomios de Chebyshev para la aproximación y del esquema de Cheon-Kim-Kim-Song (CKKS) como cifrado homomórfico. Se evalúa cómo afectan estas elecciones a la precisión del modelo y su complejidad computacional. Además, se analiza también la acumulación del componente aleatorio que se agrega normalmente en los cifrados homomórficos para proteger los datos cifrados, conocido como ruido criptográfico. Los resultados experimentales confirman que, aunque existe un ligero impacto en la precisión respecto a las funciones originales, las aproximaciones polinómicas mantienen una eficacia adecuada para aplicaciones prácticas donde la privacidad es crítica. Sin embargo, la latencia y el coste computacional actuales suponen un desafío considerable que invita a explorar técnicas criptográficas más eficientes y escalables.**

***Index Terms*—Redes neuronales profundas, Cifrado homomórfico, CKKS, Privacidad de datos, Aproximaciones polinómicas, Funciones de activación.**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

El creciente uso de redes neuronales profundas (*Deep Neural Networks*, DNNs) en aplicaciones sensibles de salud, finanzas, seguridad, etc. ha motivado la investigación en cifrado homomórfico, que permite procesar datos cifrados sin necesidad de descifrarlos [1]. Estos cifrados ofrecen una ventaja significativa al garantizar la privacidad de los datos durante su procesamiento, especialmente en escenarios donde los datos no pueden quedar expuestos, ni siquiera temporalmente. Usando este tipo de cifrado, es posible realizar inferencias y cálculos sobre redes neuronales totalmente cifradas [2]. Sin embargo, estos enfoques presentan limitaciones prácticas importantes, como un elevado coste computacional o tiempos de inferencia significativamente mayores. Además, si no se emplean esquemas criptográficos resistentes a ataques cuánticos, las redes neuronales cifradas podrían quedar expuestas ante futuros adversarios con capacidades cuánticas [3]. De hecho, el Instituto Nacional de Estándares y Tecnología (*National Institute of Standards and Technology*, NIST) ha destacado la urgente necesidad de adoptar esquemas criptográficos resistentes a ataques cuánticos para mitigar las amenazas emergentes derivadas de la computación cuántica [4].

En el presente trabajo se intenta responder a la necesidad de trasladar los procedimientos habituales en el uso de redes neuronales a un entorno seguro para proteger los datos de la potencial amenaza cuántica. Para ello se presenta aquí un estudio de las adaptaciones necesarias para las técnicas usadas en el entrenamiento e inferencia de las DNNs, y de la viabilidad en el uso de técnicas de cifrado homomórfico resistentes a ataques cuánticos.

Este trabajo se estructura de la siguiente manera. La siguiente sección, de preliminares (Sección II), introduce el contexto teórico del problema y los conceptos esenciales. La metodología (Sección III) describe las técnicas usadas en este trabajo. A continuación, los resultados (Sección IV) muestran el rendimiento de dichas técnicas, mientras que una discusión de las implicaciones de dichos resultados se incluye en la Sección V. Finalmente, la Sección VI cierra el trabajo con algunas conclusiones y trabajos futuros.

II. Preliminares

II-A. Cifrado Homomórfico

El cifrado homomórfico (*Homomorphic Encryption*, HE) es una técnica criptográfica que permite realizar operaciones directamente sobre datos cifrados, obteniendo resultados que, al ser descifrados, coinciden con los que se se habrían obtenido si las operaciones se hubieran realizado sobre los datos en claro [5]. Se llama cifrado completamente homomórfico (*Fully Homomorphic Encryption*, FHE) al HE que admite cualquier tipo de operación matemática, como sumas, productos y otras más complejas, sobre los datos cifrados. La propiedad homomórfica es especialmente útil en aplicaciones que requieren procesar información sensible sin comprometer su privacidad.

Uno de los esquemas HE más destacados es el de Cheon-Kim-Kim-Song (CKKS) [6]. Este esquema está diseñado para realizar operaciones aritméticas aproximadas sobre datos cifrados, lo que lo hace especialmente adecuado para aplicaciones que manejan números reales y requieren cálculos de precisión limitada. La característica principal del esquema CKKS es su capacidad para soportar operaciones de suma (adiciones) y producto (multiplicaciones), manteniendo un control sobre el ruido criptográfico introducido durante el proceso de cifrado y operaciones subsecuentes. Esta eficiencia homomórfica está directamente vinculada con la profundidad multiplicativa que el esquema puede tolerar antes de que el ruido criptográfico acumulado haga ininteligible el mensaje cifrado.

En el esquema CKKS, para mantener la capacidad de cómputo y renovar el cifrado cuando el nivel de ruido criptográfico se acerca a un umbral crítico, se suele recurrir al uso de *bootstrapping*[7], que realimenta los valores cifrados para restablecer la cota de ruido criptográfico sin alterar la información cifrada.

La seguridad del esquema CKKS se basa en la dificultad de una extensión del problema de aprendizaje con errores (*Learning With Errors*, LWE) [8] al contexto de los anillos polinomiales, conocido como problema de aprendizaje con errores en el anillo (*Ring Learning With Errors*, RLWE) [9]. En términos generales, el problema RLWE consiste en distinguir entre muestras obtenidas de una distribución específica que incorpora ruido criptográfico y muestras uniformemente aleatorias, dentro de un anillo de polinomios con coeficientes en un campo finito. La dificultad computacional de resolver el problema RLWE, incluso utilizando algoritmos cuánticos, proporciona una base sólida para la resistencia cuántica de esquemas de cifrado como CKKS.

II-B. Redes Neuronales

Las redes neuronales conforman un modelo computacional inspirado en la estructura y funcionamiento del cerebro humano [10]. En su forma más elemental, una red neuronal está compuesta por neuronas organizadas en capas: la capa de entrada (que recibe los datos), una o varias capas intermedias u ocultas (que llevan a cabo transformaciones progresivamente más abstractas) y la capa de salida (que produce la estimación o clasificación). Una neurona, en su versión básica, puede describirse formalmente mediante la expresión:

$$y = \phi\left(\sum_{j=1}^{n} w_j x_j + b\right),$$

donde $\phi(\cdot)$ es la función de activación, x_j son las entradas, w_j los pesos correspondientes y b el sesgo. La Figura 1 ilustra el funcionamiento de una neurona.

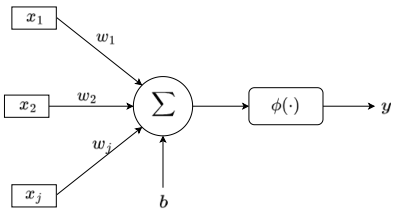


Figura 1: Esquema de una neurona artificial básica

Mediante un proceso iterativo de entrenamiento conocido como *backpropagation*, la red ajusta los parámetros (pesos y sesgos) para minimizar una función de error y, en consecuencia, mejorar su capacidad de generalización [11]. Además, gracias a su capacidad para combinar capas y funciones de activación no lineales, las redes neuronales permiten aproximar de forma arbitraria un amplio rango de funciones matemáticas, tal como se establece en el *Teorema de Aproximación Universal* [12].

En el contexto del aprendizaje automático y la ciencia de datos, las redes neuronales han demostrado su eficacia en tareas como reconocimiento de imágenes, procesamiento del lenguaje natural y predicción de series temporales [13]. Sin embargo, cuando se pretende que estas redes operen con datos cifrados, surgen limitaciones específicas que restringen tanto el proceso de entrenamiento como el de inferencia. Entre los principales retos se encuentran el coste computacional elevado y la acumulación de ruido criptográfico, especialmente en DNNs que requieren numerosas operaciones de multiplicación y funciones de activación no lineales. Aunque gran parte de la literatura se centra en el estudio del caso de la inferencia, es decir, usando un modelo ya entrenado [14], existe un creciente interés en entrenar completamente redes neuronales con datos cifrados [15] [16]. Este enfoque garantizaría la privacidad de los datos durante todo el ciclo de vida del modelo. Sin embargo, debido a la naturaleza iterativa del entrenamiento, se incrementa notablemente la sobrecarga computacional y el nivel de ruido criptográfico acumulado. Además, aproximar las funciones de activación (por ejemplo, mediante polinomios) para mantener bajo control la profundidad multiplicativa, puede afectar a la precisión del modelo. Estas dificultades hacen que, hoy por hoy, el entrenamiento de redes neuronales con cifrado homomórfico siga siendo un área activa de investigación en la que se busca reducir la complejidad y, a la vez, aprovechar las ventajas de un cifrado que preserve la privacidad de principio a fin.

Entre los autores que han comenzado a explorar la posibilidad de entrenar completamente modelos de redes neuronales con cifrado homomórfico destaca [17] como uno de los primeros intentosSin embargo, pese a optimizar la representación de los datos (fijando un número reducido de bits de precisión), y emplear técnicas de aproximación para las funciones de activación, los resultados muestran que el coste de entrenamiento sigue siendo prohibitivo para redes de gran tamaño ya que, incluso con redes y conjuntos de datos relativamente pequeños, se obtienen tiempos de cómputo que superan por varios órdenes de magnitud a la versión en claro. De hecho, ese trabajo destaca que el problema no radica solo en las múltiples operaciones de multiplicación y suma que se ejecutan en cada iteración del aprendizaje, sino también en la acumulación de ruido criptográfico y la consecuente necesidad de recurrir a técnicas de *bootstrapping*. Por tanto, estas limitaciones hacen que el entrenamiento de redes neuronales con cifrado homomórfico sea, a día de hoy, excesivamente lento para la mayoría de casos de uso industriales.

Se están explorando soluciones al problema del entrenamiento de redes neuronales con cifrado homomórfico, como el entrenamiento parcial, donde solo la fase de *fine-tuning* de ajuste de los parámetros de un modelo previamente entrenado con un conjunto grande de datos se realiza sobre datos cifrados. También se están analizando otras soluciones, como el uso de aprendizaje federado o el diseño de arquitecturas de red más ligeras, que mantengan una aceptable capacidad de clasificación o regresión. Conseguir una reducción efectiva de los requisitos computacionales y mejorar los mecanismos de preservación de la exactitud (pese a las restricciones del cifrado) constituye, por tanto, un hueco importante que sigue motivando el desarrollo de nuevas aproximaciones.

II-C. Aproximaciones Polinómicas de Funciones no Lineales

La mayoría de redes neuronales hacen uso de funciones de activación no polinómicas como tramo rectificado (*Rectified*

Linear Unit, ReLU), tangente hiperbólica o sigmoide [18]. Con HE, las multiplicaciones y sumas se pueden manejar en las redes neuronales de forma relativamente directa. Sin embargo, operaciones como la comparación, la exponencial o la ReLU resultan más costosas o directamente inviables sin utilizar *bootstrapping* intensivo. Por ello, se recurre a aproximaciones polinómicas que, en intervalos acotados, se comportan de forma similar a la activación deseada [19]. Entre las aproximaciones polinómicas más relevantes destacan:

- Polinomios de Taylor o Chebyshev. Sirven para aproximar funciones como $\mathrm{ReLU}(x) = \max(0, x)$ o $\sigma(x) = \frac{1}{1+e^{-x}}$ mediante sumas y multiplicaciones de grado bajo.
- Aproximaciones de activaciones combinadas. En ciertos casos se define una "activación polinómica" con un grado mínimo que capte la no linealidad requerida, permitiendo a la red conservar parte de su capacidad de representación.

El uso de polinomios de bajo grado disminuye la profundidad multiplicativa necesaria en el circuito homomórfico y, con ello, reduce la acumulación de ruido criptográfico o la frecuencia de *bootstrapping*. Sin embargo, la calidad de la aproximación puede impactar negativamente la convergencia y exactitud del modelo. Por ello, diseñar la mejor aproximación polinómica para cada activación y rango de valores sigue siendo un tema de investigación abierto, pues no existe una única aproximación que funcione de forma óptima en todos los contextos. De ahí que combinar técnicas criptográficas (ruido, niveles de cifrado, etc.) con criterios de optimización (arquitectura de la red, entrenamiento en claro vs. cifrado) sea un elemento clave para llevar la inferencia y potencialmente el entrenamiento a escenarios cada vez más ambiciosos.

II-D. Redes neuronales con FHE

Para que el entrenamiento y la inferencia de redes neuronales con FHE sean viables en la práctica, deben considerarse los siguientes aspectos técnicos [20]:

1. Esquema FHE con suficiente profundidad multiplicativa. Debe emplearse un esquema de cifrado homomórfico capaz de evaluar funciones complejas sobre datos cifrados sin necesidad de descifrarlos. Esto requiere una profundidad multiplicativa adecuada, que permita realizar un número suficiente de multiplicaciones consecutivas antes de que el ruido criptográfico acumulado invalide el resultado cifrado.
2. Control periódico del ruido criptográfico. Es esencial controlar o eliminar periódicamente el ruido criptográfico mediante técnicas como el *bootstrapping*, para que los resultados cifrados sigan siendo legibles después de múltiples operaciones homomórficas.
3. Infraestructura de alto rendimiento. Debido al elevado coste computacional del cifrado homomórfico, es crucial emplear hardware optimizado (por ejemplo, GPUs o aceleradores específicos para FHE) y recursos computacionales de alto rendimiento que reduzcan significativamente los tiempos de ejecución.
4. Adaptación de algoritmos de optimización. Es necesario adaptar específicamente los algoritmos de optimización empleados en el entrenamiento de redes neuronales para que operen exclusivamente mediante sumas y multiplicaciones, ya que son las únicas operaciones soportadas nativamente por los esquemas FHE.
5. Aproximación polinómica de funciones no lineales. Dado que las funciones de activación no lineales son críticas para capturar relaciones complejas en DNNs, estas deben aproximarse mediante polinomios de bajo grado, debido a que los esquemas FHE solo soportan directamente sumas y multiplicaciones.

III. Metodología

El propósito central de este trabajo es evaluar, en un entorno de cifrado homomórfico, diversas aproximaciones polinómicas para las funciones de activación empleadas en redes neuronales. En particular, se busca analizar de qué manera estas aproximaciones polinómicas afectan a:

1. La exactitud de la red, comparada con la obtenida al utilizar la función de activación original.
2. El coste computacional, que incluye el número de operaciones de multiplicación, el tiempo de ejecución y la profundidad multiplicativa necesaria en el esquema de cifrado.
3. El nivel de ruido criptográfico acumulado, evaluando si es imprescindible llevar a cabo procesos de *bootstrapping* para "refrescar" el cifrado.

Como objetivo secundario, se pretende exhibir un flujo de trabajo de inferencia cifrada sobre datos reales, empleando la mejor de las aproximaciones evaluadas, a fin de demostrar la viabilidad de estos métodos en un escenario práctico. Con ello, se obtiene un panorama tanto de la eficacia de la aproximación escogida como de la sobrecarga computacional asociada a la inferencia bajo cifrado homomórfico.

III-A. Aproximación con Polinomios de Chebyshev

En las aproximaciones realizadas, se optó por utilizar polinomios de Chebyshev debido a sus conocidas propiedades de minimización del error en el intervalo seleccionado. En particular, se utilizó para aproximar cuatro funciones de activación ampliamente usadas en redes neuronales (ReLU, tangente hiperbólica, softplus y sigmoide). A continuación, se detallan los resultados obtenidos para cada una de estas funciones, examinando visualmente la precisión de las aproximaciones con respecto a las funciones originales.

La Figura 2 muestra una comparación de cada función de aproximación usando el método de Chebyshev respecto a la original. En esos gráficos se observa cómo, al incrementar el grado del polinomio de aproximación, la función aproximada se ajusta con mayor fidelidad a la función de activación original en el intervalo considerado. En particular, para los polinomios de grado 9 se aprecia que la curva resultante prácticamente coincide con la función real en todos los casos (ReLU, Sigmoid, Softplus y Tanh). Las gráficas de la derecha muestran el error cuadrático medio (*Mean Squared Error*, MSE), que cuantifica la diferencia entre la aproximación y la función original. Tal como era esperable, cuando aumenta el grado del polinomio disminuye el error, quedando patente que el polinomio de grado 9 proporciona la menor desviación. Además, se aprecian "campanas" u oscilaciones en la forma del error, debido a la propia naturaleza de los polinomios de

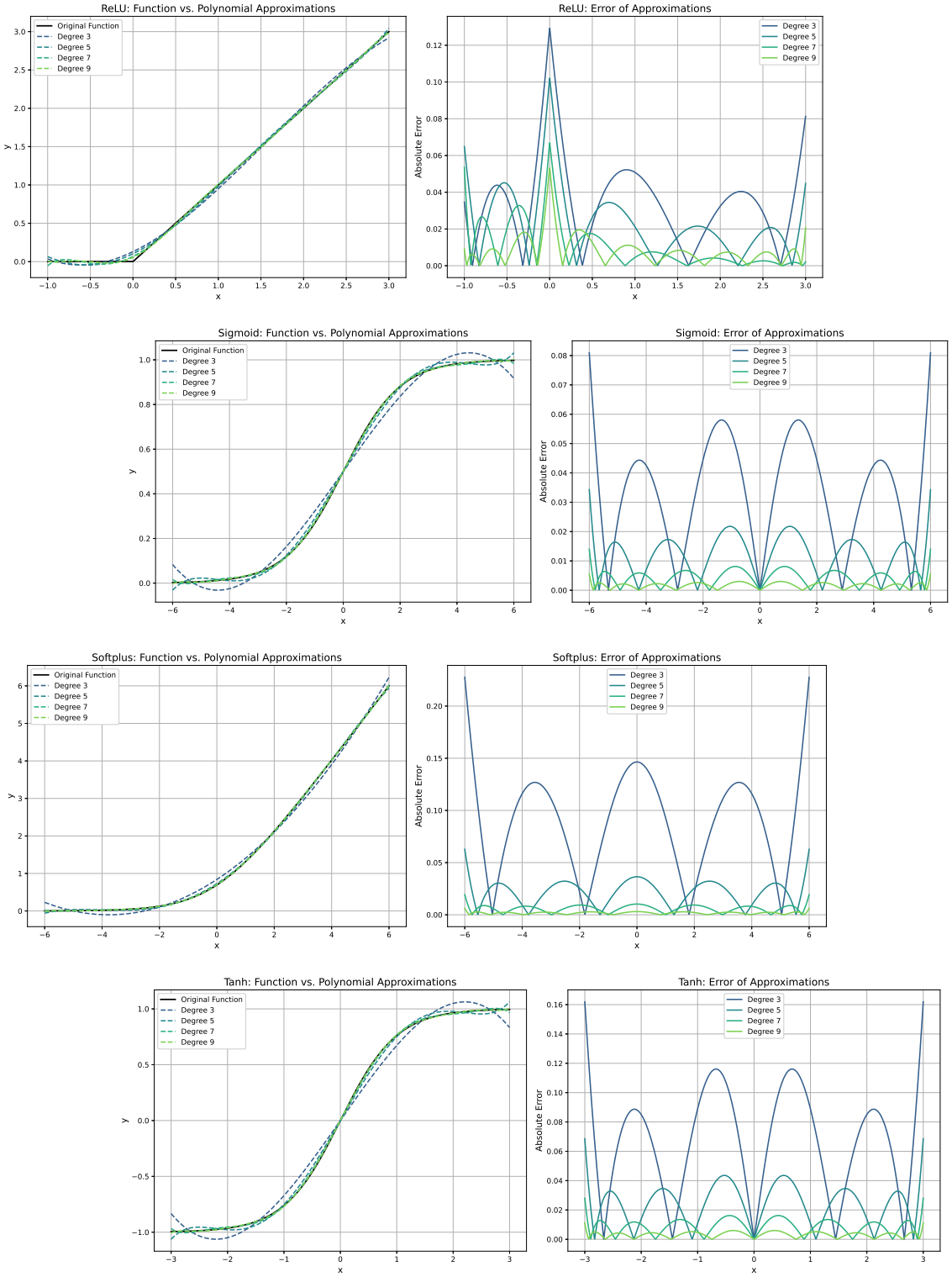


Figura 2: Comparación de cada función de aproximación usando el método de Chebyshev respecto a la original

Chebyshev, que introducen estos patrones periódicos en la aproximación sobre el intervalo de trabajo.

Por último, desde la perspectiva de los esquemas de cifrado homomórfico, es importante sopesar el grado del polinomio. Un grado mayor implica un número más elevado de coeficientes y, en consecuencia, un mayor número de operaciones de multiplicación durante la inferencia. Esto conlleva un aumento en el ruido criptográfico acumulado en el cifrado y puede exigir técnicas adicionales como el *bootstrapping* para mantener la precisión requerida.

III-B. Entrenamiento con Texto en Claro

Entrenar con las técnicas actuales una red neuronal completamente cifrada resulta poco factible debido a la limitada escalabilidad de los esquemas de cifrado. Incluso en ejemplos de complejidad reducida, los tiempos de entrenamiento pueden extenderse hasta más de 28 días, además de requerir un gran volumen de recursos computacionales y de almacenamiento. Por esta razón, una práctica habitual es realizar el entrenamiento con datos no cifrados, en claro, y, posteriormente, llevar a cabo la fase de inferencia sobre datos cifrados.

En este trabajo se utilizó el conjunto de datos MNIST [21], consistente en 60,000 imágenes de entrenamiento y 10,000 de prueba de dígitos manuscritos (0 al 9), cada una con un tamaño de 28×28 píxeles en escala de grises. Se diseñó una red neuronal sencilla que cuenta con una capa de entrada y dos capas ocultas, en las cuales se puede modificar la función de activación (ej., ReLU, Sigmoid, Softplus, Tanh). La última capa, encargada de la clasificación en 10 clases (una para cada dígito), emplea una aproximación polinómica de la función *softmax* (ver Figura 3). De esta manera, se obtiene una arquitectura suficientemente flexible para evaluar distintas aproximaciones polinómicas sin incurrir en la complejidad de DNNs de gran escala.

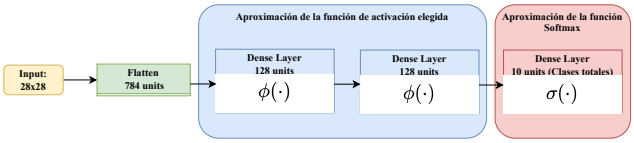


Figura 3: Aproximaciones de las funciones no lineales

Una vez completado el entrenamiento del modelo en claro, se procedió a evaluar su desempeño realizando la inferencia sobre datos cifrados utilizando el esquema CKKS. Para ello, se cifraron las muestras de entrada con la clave pública y se ejecutaron todas las operaciones de la red neuronal con los polinomios de aproximación en el dominio cifrado. A modo de comparación, también se midió el tiempo de inferencia en texto plano con el mismo modelo.

IV. Resultados

En esta sección se exponen los resultados obtenidos a lo largo de los experimentos diseñados para evaluar la viabilidad y eficacia de la aproximación polinómica de funciones de activación en un entorno de cifrado homomórfico. En primer lugar, se compara la exactitud (precisión de clasificación) alcanzada por el modelo al emplear las funciones aproximadas frente al modelo original propuesto en la Sección III-B. Posteriormente, se estudia el rendimiento de la inferencia cifrada bajo el esquema CKKS en contraste con la inferencia en claro, ilustrando tanto los tiempos de ejecución como el sobrecoste computacional asociado al cifrado. Por último, se presenta un análisis teórico que examina la acumulación de ruido criptográfico en el cifrado homomórfico cuando no se recurre a procesos de *bootstrapping*, evidenciando las posibles limitaciones y compromisos que conlleva la ejecución de entrenamiento sin "refrescar" el cifrado.

Los experimentos se llevaron a cabo en un una máquina por un procesador Intel Core i7-11700KF a 3.60GHz, en una GPU NVIDIA RTX 4070, 32 GB de memoria RAM y el sistema operativo Ubuntu 24.04.

IV-A. Impacto de la Aproximación en la Precisión

En primer lugar, se analizó la elección de los polinomios de aproximación para cada función de activación Sección III-A. Tras evaluar distintos grados, se determinó que el grado 5 ofrecía el mejor equilibrio entre precisión y complejidad computacional, dada la relación directa entre el número de coeficientes polinomiales y el coste del cifrado homomórfico. En concreto, un grado más elevado permitiría una aproximación más precisa pero incrementaría el número de operaciones necesarias, aumentando de forma notable la profundidad multiplicativa y el tiempo de ejecución durante la inferencia.

Se estudió el comportamiento del modelo cuando clasifica datos cifrados frente a su equivalente en claro (texto plano). Como puede observarse en la Figura 4, la precisión de las funciones aproximadas (barras verdes) se encuentra muy próxima a la de las versiones originales (barras azules), con diferencias que apenas llegan a unas centésimas en los valores de exactitud. Esto se aprecia en las cuatro funciones evaluadas:

- Sigmoid: se pasa de un 0.959 de precisión en la aproximación polinómica a un 0.967 con la activación original.
- Tanh: la aproximación alcanza 0.964 frente a 0.973, conservando así un nivel de exactitud elevado.
- ReLU: con la aproximación se registra 0.961, mientras que el modelo original consigue 0.973.
- Softplus: el valor con polinomios es de 0.956, comparado con 0.967 de la función original.

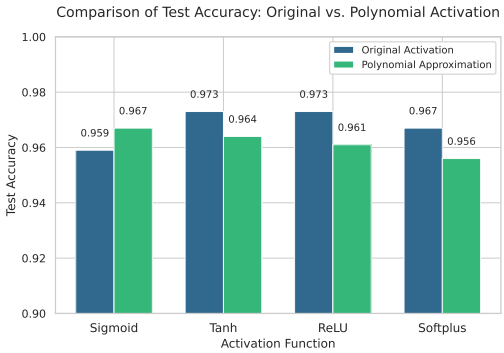


Figura 4: Precisión de funciones no lineales vs. aproximadas

A pesar de que en algunas funciones se hay pequeñas brechas de rendimiento, el margen entre la función original y la aproximación polinómica es lo suficientemente reducido como para garantizar que no se sacrifica la exactitud de manera significativa. Por ello, estos resultados confirman la viabilidad de utilizar aproximaciones polinómicas en redes neuronales que operan sobre datos cifrados, manteniendo un balance adecuado entre rendimiento y protección de la información. Este hallazgo es especialmente relevante en escenarios donde la privacidad reviste una importancia capital, dado que demuestra la posibilidad de preservar la precisión sin exponer los datos en ningún punto de la inferencia.

IV-B. Ruido Criptográfico y Eficiencia Computacional

La Figura 5 ilustra cómo crece el error acumulado en escala logarítmica para el esquema CKKS, a medida que aumenta la profundidad multiplicativa, representada en el eje horizontal como el número de operaciones de multiplicación. Cabe destacar que el eje vertical utiliza una escala logarítmica, lo cual permite apreciar que, a partir de unas 100 multiplicaciones, el incremento del error se vuelve exponencialmente más pronunciado. Esto refleja la propia naturaleza aproximada de CKKS, en la que cada operación homomórfica añade un pequeño nivel de ruido criptográfico al criptograma.

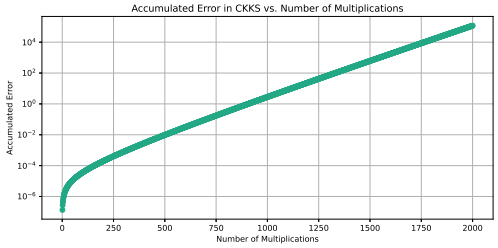


Figura 5: Error acumulado según la profundidad multiplicativa

En este caso, la red neuronal tiene pocos parámetros y la profundidad multiplicativa durante la inferencia es moderada, de modo que el umbral crítico de ruido criptográfico no se alcanza. Aun así, en aplicaciones donde se requieran centenares o miles de operaciones de multiplicación, la posibilidad de incurrir en errores de mayor magnitud demanda implementar técnicas de refresco del cifrado para garantizar un funcionamiento fiable del sistema.

La Figura 6 ilustra cómo crece el tiempo de inferencia para el texto cifrado comparado con el caso del texto plano.

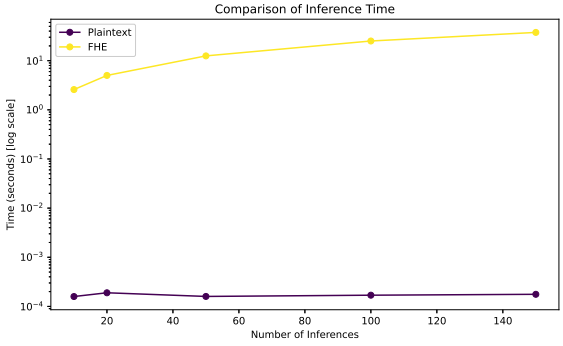


Figura 6: Tiempo de inferencia para texto cifrado vs. plano

Como era de esperar, la inferencia sobre datos cifrados presenta un tiempo de cómputo considerablemente mayor que la ejecución equivalente en claro. Esto se debe a que las operaciones aritméticas homomórficas (suma y multiplicación) conllevan un sobrecoste adicional en comparación con la aritmética tradicional en punto flotante o entero. Así, factores como la necesidad de mantener la exactitud de la aproximación, la gestión del ruido criptográfico y la profundidad multiplicativa hacen que cada operación requiera más recursos de cómputo y, por ello, se incremente la latencia.

V. Discusión

En este estudio se ha explorado la protección de la privacidad en DNNs mediante el uso de cifrado homomórfico. Se ha comprobado que, como aproximación inicial, esta técnica ofrece resultados prometedores en términos de exactitud. Sin embargo, el sobrecoste en latencia y recursos computacionales puede ser prohibitivo en escenarios que requieren un rendimiento muy elevado. A pesar de ello, en aplicaciones críticas o altamente sensibles, el compromiso entre eficiencia y privacidad puede justificarse, dada la importancia de preservar la confidencialidad de los datos en todo momento.

Asimismo, los experimentos han mostrado que, con los esquemas de cifrado y flujos de trabajo actuales, llevar a cabo un entrenamiento plenamente cifrado en redes neuronales de mayor envergadura resulta inviable, debido tanto a la ralentización extrema de los cálculos como a la acumulación significativa de error. En cambio, la aproximación de entrenar en claro y ejecutar la fase de inferencia bajo cifrado homomórfico se presenta como una alternativa viable cuando la privacidad del modelo y/o de los datos es prioritaria.

Además, se ha observado que el uso de activaciones polinómicas mantiene un nivel de precisión cercano al de las funciones originales, lo que evidencia que la pérdida de exactitud es mínima cuando se emplean métodos compatibles con la aritmética homomórfica. Este hallazgo sugiere que, aunque se requiere un mayor esfuerzo en la implementación de operaciones polinómicas y en la gestión del ruido criptográfico, el enfoque es válido en escenarios donde la integridad y la confidencialidad de la información son fundamentales.

VI. Conclusiones

En este trabajo se ha investigado cómo proteger la privacidad en redes neuronales profundas utilizando cifrado homomórfico. Concretamente, se ha demostrado que la adopción de funciones de activación polinómicas resulta ser una buena solución para superar las limitaciones operativas del cifrado homomórfico, manteniendo una precisión similar a la obtenida con funciones no lineales estándar. Este enfoque indica que existen métodos compatibles con la aritmética homomórfica que permiten preservar la utilidad del modelo, incluso en contextos de protección de la privacidad.

Una posible orientación de la investigación futura de esta línea de trabajo es hacia el desarrollo de técnicas más eficientes y escalables. Se hace necesario el diseño de esquemas criptográficos optimizados que reduzcan la latencia y la necesidad de procesos de *bootstrapping*. Mejorar estas dimensiones será crucial para acercarse a redes neuronales completamente cifradas que puedan abordar problemas de mayor complejidad sin comprometer el rendimiento.

Agradecimientos

Esta investigación ha sido posible gracias a los proyectos PID2022-138933OB-I00 ATQUE y 2023DIG28 IACTA financiados por MCIN/AEI/10.13039/501100011033 y Fundaciones Cajacanarias y Caixa, y a la Cátedra de Ciberseguridad de ULL, INCIBE y la Unión Europea NextGenerationEU/PRTR.

REFERENCIAS

[1] A. Wood, K. Najarian, and D. Kahrobaei, "Homomorphic encryption for machine learning in medicine and bioinformatics," *ACM Comput. Surv.*, vol. 53, no. 4, 2020.

[2] J.-W. Lee, H. Kang, Y. Lee, W. Choi, J. Eom, M. Deryabin, E. Lee, J. Lee, D. Yoo, Y.-S. Kim, and J.-S. No, "Privacy-preserving machine learning with fully homomorphic encryption for deep neural network," *IEEE Access*, vol. 10, pp. 30 039–30 054, 2022.

[3] K. Han, W.-K. Lee, A. Karmakar, M.-K. Yi, and S. O. Hwang, "Quripfe-net: Quantum-resistant ipfe-based neural network," *IEEE Transactions on Emerging Topics in Computing*, 2024.

[4] N. I. of Standards and Technology, "Post-quantum cryptography," https://www.nist.gov/programs-projects/post-quantum-cryptography, accessed: 2025-02-16.

[5] C. Marcolla, V. Sucasas, M. Manzano, R. Bassoli, F. H. P. Fitzek, and N. Aaraj, "Survey on fully homomorphic encryption, theory, and applications," *Proceedings of the IEEE*, vol. 110, no. 10, pp. 1572–1609, 2022.

[6] J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *Advances in Cryptology – ASIACRYPT 2017*, T. Takagi and T. Peyrin, Eds. Cham: Springer International Publishing, 2017, pp. 409–437.

[7] C. Gentry, S. Halevi, and N. P. Smart, "Better bootstrapping in fully homomorphic encryption," Cryptology ePrint Archive, Paper 2011/680, 2011. [Online]. Available: https://eprint.iacr.org/2011/680

[8] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," *J. ACM*, vol. 56, no. 6, Sep. 2009. [Online]. Available: https://doi.org/10.1145/1568318.1568324

[9] V. Lyubashevsky, C. Peikert, and O. Regev, "On ideal lattices and learning with errors over rings," *J. ACM*, vol. 60, no. 6, Nov. 2013. [Online]. Available: https://doi.org/10.1145/2535925

[10] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[11] C. Rohlfs, "Generalization in neural networks: A broad survey," 2024. [Online]. Available: https://arxiv.org/abs/2209.01610

[12] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0893608089900208

[13] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, p. e00938, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844018332067

[14] A. Stoian, J. Frery, R. Bredehoft, L. Montero, C. Kherfallah, and B. Chevallier-Mames, "Deep neural networks for encrypted inference with TFHE," Cryptology ePrint Archive, Paper 2023/257, 2023. [Online]. Available: https://eprint.iacr.org/2023/257

[15] L. Colombo, A. Falcetta, and M. Roveri, "Training encrypted neural networks on encrypted data with fully homomorphic encryption," in *Proceedings of the 12th Workshop on Encrypted Computing & Applied Homomorphic Cryptography*, ser. WAHC '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 64–75. [Online]. Available: https://doi.org/10.1145/3689945.3694802

[16] L. Montero, J. Frery, C. Kherfallah, R. Bredehoft, and A. Stoian, "Neural network training on encrypted data with tfhe," 2024. [Online]. Available: https://arxiv.org/abs/2401.16136

[17] K. Nandakumar, N. Ratha, S. Pankanti, and S. Halevi, "Towards deep neural network training on encrypted data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[19] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 201–210. [Online]. Available: https://proceedings.mlr.press/v48/gilad-bachrach16.html

[20] C. Liu, Z. Li, and X. Li, "Accelerating fully homomorphic encryption: A survey on hardware and software solutions," *IEEE Access*, vol. 9, pp. 123 456–123 470, 2021.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

Diseño de un sistema de votación electrónica basado en Blockchain y cifrado homomórfico

Javier Padilla Pío
Universidad de La Laguna
San Cristóbal de La Laguna, 38203
jpadillp@ull.edu.es

Cándido Caballero-Gil
Universidad de La Laguna
San Cristóbal de La Laguna, 38203
ccabgil@ull.edu.es

Jezabel Molina-Gil
Universidad de La Laguna
San Cristóbal de La Laguna, 38203
jmmolina@ull.edu.es

Moti Yung
Universidad de Columbia
New York, 10027
motiyung@gmail.com

***Resumen*—El objetivo principal de este trabajo es desarrollar un protocolo de votación electrónica que aproveche las características del Blockchain y la criptografía homomórfica para garantizar un proceso seguro y transparente. Un sistema de votación eficaz debe asegurar el anonimato del votante y la integridad del sistema, de manera que se preserve la confidencialidad del voto individual mientras se garantiza que cada voto sea contabilizado correctamente. Asimismo, es fundamental que el resultado de la votación permanezca desconocido hasta el momento del escrutinio para evitar influencias indebidas en el proceso. Para cumplir con estas condiciones, en nuestro sistema todos los votos son cifrados en el cliente, asegurando que el servidor no pueda vincular el voto con el votante, ya sea en su forma cifrada o descifrada. Similar a los sistemas tradicionales, los usuarios depositarán sus votos cifrados en una urna digital, que en este caso sería una cadena de bloques. Así, el almacenamiento y el posterior procesamiento de los votos para el recuento serán inmutables y transparentes para los usuarios.**

***Index Terms*—Blockchain, cifrado homomórfico, ElGamal, RSA, BFV, ZKP**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Los procesos electorales son pilares fundamentales en la configuración de gobiernos y organizaciones democráticas. Tradicionalmente, se realizan de manera presencial, utilizando papeletas y urnas físicas. Si bien este método ha demostrado ser efectivo, enfrenta diversos desafíos, como costes elevados, riesgo de errores humanos y fraude electoral. Esto impulsa la búsqueda de alternativas más eficientes, seguras y confiables. En este sentido, el voto electrónico surge como una opción prometedora.

Además, los métodos tradicionales pueden limitar significativamente la participación de ciertos grupos de la población, como personas con discapacidad, residentes en zonas rurales, ciudadanos que trabajan en el extranjero o que cumplen misiones diplomáticas. Esta situación ha impulsado la búsqueda de métodos alternativos, más accesibles y eficientes, como la implementación del voto electrónico, la ampliación del voto por correo o la creación de centros de votación móviles. Garantizar el acceso universal al voto es fundamental para fortalecer las democracias y construir sociedades más justas e inclusivas.

En este contexto, los sistemas de votación electrónica surgen como una solución prometedora ya que tienen el potencial de simplificar y agilizar el proceso electoral, reduciendo costes y minimizando errores. Además, un sistema de votación en línea podría mejorar significativamente la accesibilidad, permitiendo que un mayor número de ciudadanos participe en las elecciones con mayor comodidad y sin las restricciones inherentes a los métodos tradicionales.

Aunque los sistemas de votación electrónica ofrecen numerosos beneficios, su implementación también presenta varios desafíos críticos que deben ser abordados para asegurar su eficacia y seguridad. Existen numerosas investigaciones que abordan los principales problemas que presentan los sistemas de votación a través de internet. En el artículo “Going from bad to worse: from Internet voting to Blockchain voting” [1] se presentan una serie de cuestiones que deben ser resueltas de cara a la implementación de un sistema a nivel nacional.

En este trabajo no se pretende dar respuesta a todas las preguntas planteadas en el artículo, puesto que muchas se basan en políticas de seguridad independientes de los aspectos tecnológicos. sino a los aspectos más técnicos como los principales objetivos de este trabajo:

- **Seguridad y vulnerabilidades**. La dependencia en la tecnología introduce riesgos significativos de ciberataques, tanto para la obtención de información confidencial como para causar disrupciones en el sistema. Es esencial implementar medidas robustas de seguridad para proteger el sistema contra posibles vulnerabilidades y garantizar su resiliencia.
- **Privacidad y confidencialidad**. La privacidad del votante debe ser preservada a toda costa, asegurando que ninguna entidad, incluida la administración electoral, pueda conocer el contenido de los votos individuales. Esto es fundamental para mantener la confianza en el proceso electoral.
- **Transparencia**. Todo el proceso de votación debe ser transparente y verificable, permitiendo que tanto los votantes como entidades externas de confianza puedan auditar el sistema. La transparencia es crucial para garantizar la integridad y la legitimidad del proceso electoral.
- **Accesibilidad**. Aunque la votación electrónica puede facilitar el proceso para una parte de la población, existe

el riesgo de excluir a aquellos sin acceso a la tecnología necesaria o sin los conocimientos técnicos para utilizarla. Es fundamental diseñar sistemas inclusivos que aseguren la participación de todos los ciudadanos.

A pesar de estos desafíos, resolver cada uno de los puntos mencionados es crucial para obtener el respaldo de la sociedad. La aceptación pública no se puede simplemente codificar en un protocolo; es el resultado de un esfuerzo continuo para identificar y mitigar todos los posibles problemas asociados con un nuevo sistema de votación electrónica. Sin esta confianza, la adopción de dichos sistemas será limitada y su legitimidad cuestionada.

II. Contexto teórico

II-A. Cifrados homomórficos

Los criptosistemas homomórficos son aquellos que, dada una función algebraica $f(x, y)$, y una función de cifrado $C(m)$ cumplen que $C(f(x, y)) = f(C(x), C(y))$. Esta característica resulta muy interesante para este trabajo, puesto que abre las puertas a poder sumar todas las papeletas sin tener que descifrarlas.

Aunque existen numerosos criptosistemas homomórficos, se ha escogido trabajar con ElGamal [2]. Originalmente se trata de un sistema homomórfico para la multiplicación, aunque aprovechando las propiedades de las potencias se puede desarrollar un sistema homomórfico para la suma. Además, como se puede ver en el trabajo de Ronghao Zhou y Zijing Lin [3], la exponenciación del mensaje no solo cambiaría la operación homomórfica, sino que añadiría un grado de seguridad extra.

II-B. Firma ciega

En Fig. 1 se trata el protocolo de firma ciega que permite a un usuario obtener un mensaje firmado por una entidad sin que esta vea el contenido del mensaje. Se puede lograr aprovechando las propiedades de la aritmética modular presentes en RSA [4].

Este protocolo se divide en tres algoritmos o fases entre dos partes: el usuario y el firmante.

- **Configuración inicial.** Se generan los parámetros del criptosistema, así como el par de claves a utilizar.
- **Cegado.** El usuario genera un mensaje m y aplica una función de cegado usando un factor aleatorio r. Luego lo envía al firmante.
- **Firma.** El firmante aplica su algoritmo de firma usando su clave privada.
- **Descegado.** El usuario elimina el efecto del cegado utilizando r para obtener una firma válida sobre el mensaje original.
- Validación. Cualquier usuario puede verificar la firma utilizando la clave pública.

Dado un número aleatorio r se cumple que $r^{ed} \equiv r \mod N$, siendo e, d, N los parámetros del criptosistema. Esta característica se aprovecha para "cegar" el mensaje a firmar multiplicándolo r elevado a la clave pública. Una vez firmado por el servidor, el usuario podrá extraer r dividiendo por este.

$\mathcal{U}suario$		$\mathcal{S}ervidor$
$m' = mr^e \mod N$	$\xrightarrow{m'}$	
	$\xleftarrow{s'}$	$s' = (m')^d \mod N$
$s = s'r^{-1} = m^d \mod N$		

Figura 1. Protocolo de firma ciega

II-C. ZKP

Las pruebas de conocimiento nulo permiten demostrar el conocimiento acerca de un secreto sin desvelar ninguna información acerca de él [5]. Toda prueba válida debe cumplir tres propiedades:

- **Completitud.** Si la afirmación es verdadera y ambas partes siguen el protocolo correctamente, entonces el verificador siempre aceptará la prueba.
- **Solidez.** Si la afirmación es falsa, un probador malicioso no podrá convencer al verificador de que la afirmación es verdadera.
- **Conocimiento nulo.** Si la afirmación es verdadera, el verificador no obtiene ninguna información adicional sobre la afirmación más allá de su veracidad.

Una aplicación de este concepto es demostrar la validez de un voto, entendido como $v = \{v_i \mid v_i \in \{0, 1\} \wedge \exists! v_i : v_i = 1\}_0^{n-1}$. Este problema se puede abordar dividiéndolo en dos más simples: verificar que todas las componentes del vector están en el conjunto $\{0, 1\}$ y que solo existe un 1. En este trabajo se utilizarán las firmas de anillo de Borromeo [6] junto a la heurística de Fiat-Shamir [7] que permite probar el conocimiento sin necesidad de interactuar con el verificador.

II-D. Blockchain

Una cadena de bloques o Blockchain representa un sistema de almacenamiento en el que los escritores publican información para todos los lectores. Estas escrituras se realizan de tal manera que no se puedan modificar ni borrar los datos publicados previamente. Partiendo de esta idea de almacenamiento inmutable y transparente para el usuario, se han desarrollado sistemas como Ethereum [8] que presentan un sistema de cómputo dentro de la cadena. Así es que se puede ejecutar código definido por cualquier usuario a través de los llamados contratos inteligentes.

Un sistema de votación puede aprovechar esta transparencia para ejecutar código y almacenar los datos como una urna en la que cada usuario deposite su voto directamente.

Otra característica fundamental de la cadena de bloques es su naturaleza distribuida. La cadena está formada por multitud de nodos, cada uno con una copia exacta de la cadena que, a través de un mecanismo de consenso, añaden nuevos bloques. En caso de existir un fallo en uno de los nodos, el resto podría suplir la carga sin perder información. El mecanismo de consenso, además, garantiza la penalización a los nodos deshonestos, evitando que añadan bloques fraudulentos. Un nodo deshonesto tendría que tener el control del 51 % [9] de la red para poder emitir bloques sin necesidad de aprobación por parte del resto de nodos.

III. Protocolo propuesto

El protocolo de votación se fundamenta en el principio de la separación del voto cifrado del usuario que lo emite. Esta medida asegura que, incluso con acceso a la clave de cifrado, resulte imposible asociar el contenido de una papeleta con un usuario específico.

Para lograr este objetivo, el proceso de votación implica la interacción entre tres entidades: el votante, el servidor de autenticación y el servidor de almacenamiento público (Blockchain). Cuando un usuario desea participar en una elección, primero debe autenticarse con el servidor de autenticación. Una vez autorizado, se le proporcionará un tique que permite el acceso al almacenamiento público.

El servidor de autenticación almacenará una serie de metadatos relacionados con cada votación, como los tiques emitidos y ciertas credenciales de recuperación de los usuarios.

La clave de este proceso reside en que el tique no puede vincularse a la cuenta del usuario. Esta condición garantiza que el servidor de autenticación no es capaz de enlazar el voto emitido con el usuario que solicitó el tique.

En esta sección se detallará cómo se ha implementado este protocolo desde una perspectiva teórica (Figura 2).

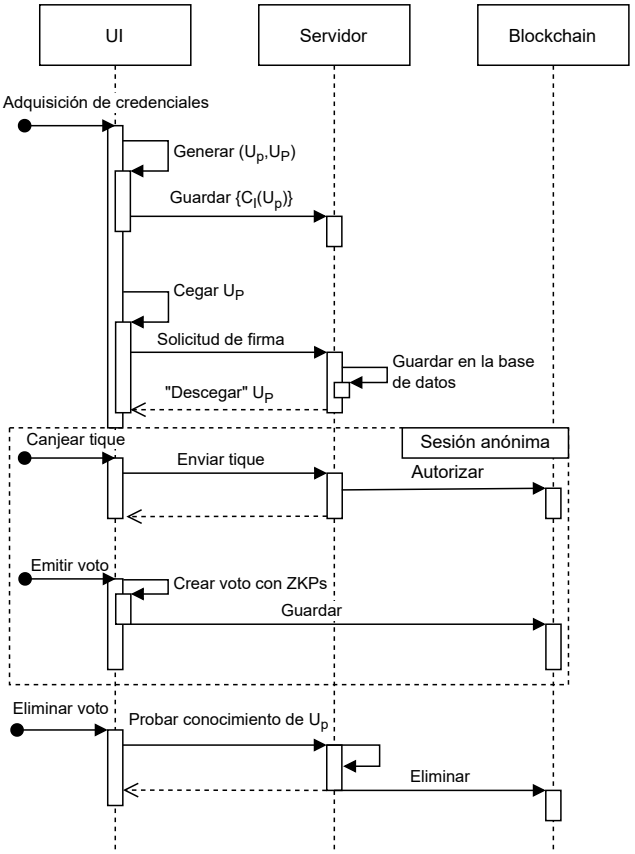


Figura 2. Acciones del usuario

III-A. Notación

Se denota un par de claves como (X_P, X_p), donde X denota qué clave se está utilizando y el subíndice si se trata de la componente pública P o privada p. Siguiendo con esta notación, la función de cifrado y descifrado, que utiliza la clave pública y privada respectivamente, se representa como $C_X(m), D_X(m)$. Además, se utilizará $S_X(m)$ para denotar una función de firma utilizando la clave privada. A continuación, se describen las primitivas utilizadas a lo largo del trabajo.

- G_p: Clave privada de ElGamal. Se crea al principio de la votación dividida en k partes, siendo k menor al número de administradores (se detalla en Sección III-C). No es hasta que acabe el proceso electoral que todos los administradores presentaran sus claves para así poder recuperar la clave privada original.
- G_P: Clave pública de ElGamal. Se crea al principio de la votación y será utilizada para cifrar todos los votos emitidos. Se utilizará una variante exponencial del cifrado para poder operar de manera homomórfica para la suma. Además, como medida extra de seguridad el criptositema se basará en la complejidad del logaritmo discreto sobre una curva elíptica.
- (R_p, R_P): Par de claves RSA. Se genera por el servidor para firmar los tiques enviados por los usuarios que solicitan acceso a una votación.
- (U_p, U_P): Par de claves de una cuenta Blockchain. Se genera por cada voto que se vaya a guardar en la cadena. Si un usuario decide cambiar su voto, tendrá también que cambiar este par de claves.
- (I_p, I_P): Par de claves del certificado digital del usuario. Este puede ser derivado de la biometría del usuario o extraído de su documento de identidad o similar.

Tanto el par I como U pueden corresponder a cualquier criptosistema asimétrico seguro, siempre y cuando implementen el cifrado y la firma. En este trabajo el par U corresponde a ECDSA [10], concretamente la curva secp256k1 [11] ampliamente utilizada en Blockchain. Para I se ha utilizado RSA con un tamaño de clave mayor a $2048b$, puesto que es el más distribuido a la hora de generar certificados digitales a nivel nacional. La restricción del tamaño de clave viene definida tanto por garantizar la seguridad del criptosistema, como por el tamaño de las claves privadas del criptosistema usado para las cuentas Blockchain (Sección III-D).

III-B. Restricciones

El protocolo ha sido diseñado para que cada votante pueda emitir un voto con un solo candidato seleccionado. En caso de querer emitir otro voto, el sistema obligará a eliminar el emitido previamente. El sistema será capaz de detectar cualquier voto mal formado y omitirlo en el recuento final.

III-C. Configuración criptográfica inicial

Al iniciar una elección, se procede con la creación de las credenciales necesarias que utilizará el servidor para cifrar y firmar:

- Los administradores generan un par de claves usando ElGamal exponencial, denotadas como (G_P, G_p), y envía la clave pública al servidor.
- Posteriormente, el servidor genera un par de claves RSA (R_P, R_p) específicamente para la elección en curso.

La clave privada de ElGamal se dividirá entre todos los administradores de tal manera que deba existir un consenso entre la mayoría de ellos para poder descifrar el resultado.

Esto se puede lograr utilizando el esquema de Shamir [12] con un umbral (k, n) tal que $k < n$. De esta manera se puede recuperar la clave con una mayoría de administradores y no forzando a estar todos.

Este proceso establece las bases criptográficas necesarias para asegurar la integridad y confidencialidad del sistema de votación electrónica.

III-D. Adquisición de credenciales

Antes de proceder, se presupone que el usuario posee un certificado digital válido y ha completado la autenticación con el servidor.

- En el cliente, el usuario genera (U_p, U_P) que se almacenarán en su dispositivo.
- El usuario calcula $C_I(U_p)$ y lo guarda en el servidor de autenticación. Este procedimiento permite al usuario recuperar sus credenciales de la cadena de bloques en caso de pérdida de la copia local.

III-D1. Solicitar tique: El acceso a la cadena de bloques está controlado por el servidor de autenticación. Controla tanto el gas necesario para ejecutar los contratos como la propia autorización interna del contrato. En esta sección se describe el procedimiento para solicitar el gas y la autorización al servidor sin desvelar la identidad del usuario.

- El usuario ciega U_P como se describe en la Sección II-B y lo envía al servidor.
- El servidor firma la información utilizando R_p y registra en la base de datos que se ha emitido un tique al usuario.
- El usuario elimina la parte aleatoria, obteniendo U_P correctamente firmado por el servidor.

El servidor hará pública una lista con todos los usuarios que han generado un tique para votar (Tabla I). Esta lista contendrá un campo por cada usuario con la firma del resto de registros. De esta manera, se garantiza que todos los tiques emitidos han sido solicitados por un usuario con un certificado válido.

Esta medida garantiza al usuario que no se han emitido más votos a la cadena de los solicitados por los usuarios. Para que el servidor pueda alterar esta tabla, tiene que o bien conocer las claves privadas de los certificados de los usuarios, o ser capaz de emitir certificados falsos. Esto destaca la importancia de separar la Autoridad Certificadora del servidor de autenticación.

III-E. Canjear tique

Tras realizar estos pasos, el servidor conoce la intención del usuario de acceder al almacenamiento público para depositar su voto. No obstante, no sabe con qué credenciales lo hará, garantizando así el anonimato. El servidor solo permitirá que se realice este proceso una vez por usuario, a no ser que este elimine su voto.

Si los tiques se canjean inmediatamente, el servidor podría relacionar la dirección de la cuenta Blockchain con la identidad del usuario. Es por esto que, para evitar cualquier tipo de correlación entre la solicitud del tique y el canjeo del mismo, se añadirá un tiempo de espera obligatorio (T). Este tiempo se calculará de manera aleatoria en el cliente a la hora de crear el tique y se adjuntará a la hora de firmarlo. De esta manera, cuando se canjee, si no ha pasado el tiempo necesario, la acción no se realizará.

A continuación, se debe canjear el tique para obtener la autorización y el gas. Esto se realiza a través de un portal anónimo (un dispositivo distinto, navegación sin cookies, etc.). El usuario envía $\{U_P, T, S(U_P, T)\}$ y el servidor reconoce la autorización basada en la firma, sin obtener información acerca de la identidad del usuario. Este tique se almacenará en el servidor para prevenir su reutilización.

III-F. Emisión del voto

Una vez autorizado a emitir el voto, el usuario deberá cifrar su elección y generar las ZKP asociadas para probar su validez.

- El usuario crea un voto y lo cifra utilizando G_P. El voto consiste en un vector de ceros y unos que representa las opciones afirmativa o negativa para cada candidato.

$$v = \{0, 0, 1\} \rightarrow v' = \{C_G(r_0, v_0), ...\} \qquad (1)$$

- El usuario podrá almacenar los escalares aleatorios utilizados para cifrar el voto, de tal manera que pueda verificar que la función de cifrado es correcta.
- Para garantizar la integridad y la corrección del voto durante el recuento, se generan dos ZKP: una para verificar que cada elemento del vector es 0 o 1, y otra para asegurar que solo un elemento del vector contiene un 1.
- El usuario ejecuta el contrato asociado a la elección directamente, utilizando su par de claves, y guarda el voto junto con las dos pruebas generadas.

III-G. Eliminación del voto

En caso de coacción o cambio de opinión, el sistema debe contar con la opción de eliminar un voto ya emitido.

- El usuario presenta la clave pública utilizada para cifrar el voto así como una ZKP con la que pueda demostrar el conocimiento de la componente privada.
- El servidor ejecuta un contrato para revocar el voto en la cadena. Solo el servidor puede ejecutar este contrato, ya que se verifica la dirección desde la cual se inicia la operación.
- El servidor registra esta operación, asociándola con el usuario y la cuenta eliminada. De esta manera, se tiene un registro de todas las operaciones críticas que pudieran suponer un fraude.

En caso de haber perdido las credenciales, el usuario podrá utilizar su certificado electrónico para descifrar las credenciales almacenadas en el servidor.

III-H. Recuento final

Una vez finalizada la elección, el servidor ejecuta una función del contrato electoral para verificar y sumar todos los votos cifrados dentro de la cadena. Esto se realiza utilizando contratos precompilados capaces de realizar operaciones con criptosistemas de ElGamal.

Finalmente, un administrador solicita el recuento al servidor. Para poder descifrar el resultado, k administradores se reunirán para reconstruir la clave privada y así descifrar el resultado.

Tabla I
TABLA VOTEINTENTS

Campo	Descripcción
Id	Identificador del voto.
UserId	Identificador del usuario.
ElectionId	Identificador de la elección para la que se ha emitido el voto.
RecoveryEthSecret	Clave privada de Blockchain cifrada utilizando I_P. El usuario la puede solicitar en caso de perder las credenciales.
Signature	Firma de los otros campos calculada utilizando el certificado del usuario: $S_I(\{\text{id, userId, RecoveryEthSecret}\})$. Permite que cualquier usuario verifique la validez del registro.

IV. CONSIDERACIONES DE SEGURIDAD

El objetivo del sistema es que el usuario no tenga que confiar en él, sino que tenga un nivel de transparencia tal que se pueda ver todo el proceso. Esto implicaría que la administración de las elecciones tendría que ser totalmente corrupta para poder modificar el resultado.

Por una parte, tenemos el almacenamiento público: totalmente distribuido, redundante, inmutable y transparente. Como se ha mencionado anteriormente, para poder actuar de manera deshonesta se debería tener el control del 51 % de la red, algo que a gran escala resulta inviable. En el peor de los casos en que la cadena de bloques se viera comprometida, el atacante tan solo podría eliminar votos emitidos y sustituirlos por otros emitidos con credenciales distintas. Esto se podría detectar fácilmente por los usuarios al revisar si su voto ha sido contabilizado correctamente, puesto que no lo encontrarían en el resultado final. En ningún caso el atacante sería capaz de emitir más votos de los emitidos de manera honesta. Esto se debe a que el servidor de autenticación hace pública una lista con todos los usuarios y su intención de emitir un voto (han solicitado un tique). Gracias a la firma de estos registros, resulta imposible modificarlos. El sistema siempre debe cumplir que el número de tiques emitidos sea menor que el número de votos emitidos.

Por otro lado, en cuanto a la coacción del usuario, cualquier tipo de sistema de voto online facilita que un usuario sea forzado a cambiar su voto. Por ello, el sistema ofrece la opción de cambiar el voto tantas veces como el usuario quiera antes del recuento. Como ya se ha mencionado, en caso de perder las credenciales, siempre será capaz de recuperarlas, puesto que se guarda una copia cifrada con el certificado del usuario en el servidor. Este procedimiento reduce el impacto de la coacción, puesto que habría que retener al usuario hasta el final de las elecciones para que no cambie su voto.

IV-A. Transparencia

En secciones anteriores se ha mencionado que se debe guardar cierta información acerca del usuario durante el desarrollo de las elecciones. A continuación, se detalla el esquema de los conjuntos de datos que se almacenarán.

En primer lugar, cada vez que un usuario inicie la adquisición de credenciales y solicite un tique cegado, se añadirá un registro a la tabla "Vote Intents" (Tabla I).

Tabla II
TABLA USERS

Campo	Descripcción
UserId	Identificador del usuario.
Name	Nombre del usuario.
Certificate	Certificado digital del usuario, utilizado para la autenticación.
CN	"Common name" o identificador del certificado.

Tabla III
TABLA ISSUEDTIQUES

Campo	Descripcción
Id	Identificador del tique.
Address	Clave pública de Blockchain a la que autoriza el tique.
ElectionId	Identificador de la elección para la que se ha emitido el voto.
Revoked	Verdadero si el voto emitido con este tique ha sido revocado.
RevokedBy	En caso de que se borre el voto, se guarda qué usuario hizo la solicitud, para así poder elaborar una traza en caso de existir comportamientos deshonestos.

La información de esta tabla se extiende con la propia información que el servidor almacena del usuario (Tabla II) común a todas las elecciones.

Cuando un usuario envíe su tique al servidor para canjearlo y recibir la autorización para votar dentro de la Blockchain, el servidor guardará también una serie de metadatos (Tabla III).

IV-B. Amenaza cuántica

La seguridad de este sistema se basa principalmente en el problema del logaritmo discreto [13], por lo que es vulnerable a un ataque cuántico utilizando el algoritmo de Shor [14]. Como se discute en la Sección V ya existen criptosistemas homomórficos que podrían ser implementados en este protocolo. No obstante, aunque en el trabajo [15] se muestran diversas construcciones válidas para desarrollar ZKPs basándose en el problema del aprendizaje con errores, los tamaños de las pruebas y su tiempo de cómputo dificultan la implementación en un sistema a gran escala.

La cadena de bloques no solo utiliza criptografía asimétrica basada en el logaritmo discreto, sino funciones de hash. La computación cuántica también supone una amenaza, pudiendo realizar búsquedas no informadas en $O(\sqrt{2^n})$ donde el espacio de búsqueda depende de n [16].

Aunque la urgencia de migrar a sistemas postcuánticos es innegable, la transición práctica hoy resulta inviable. Como ya se ha mencionado, a pesar de los existentes estándares ya publicados como ML-KEM [17] o ML-DSA [18], las alternativas postcuánticas para ZKPs siguen atascadas en una fase incipiente: los protocolos propuestos son ineficientes y demandan recursos computacionales prohibitivos. Implementarlos prematuramente, sin optimizar su escalabilidad o garantizar interoperabilidad con sistemas existentes, introduciría riesgos mayores a los que buscan mitigar. La prioridad, por ahora, debe ser consolidar marcos teórico-prácticos que equilibren resistencia cuántica, eficiencia y usabilidad. Solo entonces podrá evaluarse su adopción sin comprometer la seguridad ni la confianza en un sistema crítico como este.

V. Análisis de resultados

Es esta sección se presentan los resultados de prueba de carga sobre el algoritmo de cifrado durante la suma de votos y el descifrado del resultado. En pruebas previas realizadas con 100 iteraciones por prueba, se pudo observar que la desviación típica era despreciable. Teniendo en cuenta esto y el elevado coste computacional, los datos mostrados en este trabajo corresponden a 10 iteraciones por prueba, ya que los datos se pueden escalar sin comprometer la validez del análisis.

Para las pruebas se ha utilizado un Intel Xeon con 12 núcleos y 24GiB de RAM. Además de comprobar el rendimiento de ElGamal, se aprovechó para comprobar el de BFV [19] también. Se trata de un sistema totalmente homomórfico diseñado para la aritmética de números enteros, basado en el problema de El Aprendizaje con Errores (LWE por sus siglas en inglés) [20] (resistente a la cuántica).

La figura 3 muestra los resultados de descifrar el resultado de una elección con distinto número de votos. Se puede apreciar cómo existe una clara diferencia de rendimiento entre los dos criptosistemas. Esta diferencia se debe a la dificultad que introduce la exponenciación de ElGamal: se debe resolver el problema del logaritmo discreto utilizando una tabla de sustitución. Por tanto, cuanto mayor sea el valor cifrado, mayor debe ser la tabla, resultando en una búsqueda más lenta.

Por otra parte, se puede apreciar la clara ventaja de BFV en cuanto a almacenamiento, puesto que a diferencia de ElGamal que cifra cada uno de los candidatos de un voto de manera independiente, este sistema permite cifrar todo el voto en un solo polinomio. Esto implica que el número de candidatos no afecta al tamaño del voto ni al tiempo de cómputo, puesto que los polinomios tienen un tamaño fijo de 2048 términos.

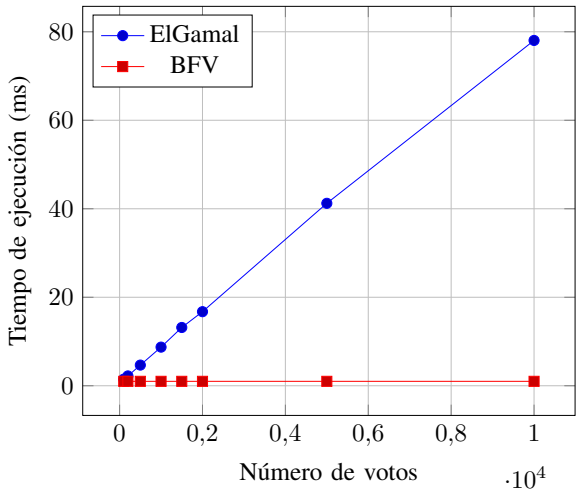


Figura 3. Tiempo de ejecución del descifrado

VI. Conclusiones y líneas futuras

El avance de la tecnología y la seguridad informática permite aproximarse a una implementación real de un protocolo de votación en línea que brinde a todos los usuarios una forma de votar de manera cómoda y segura, garantizando que su voto sea contabilizado correctamente. En este trabajo se ha estudiado una de las posibles soluciones a los retos que propone una votación segura a través de internet. Si bien su resistencia a ataques informáticos y posibles vulnerabilidades requiere de un análisis más profundo, se plantea el uso de Blockchain para facilitar la confianza de manera distribuida en el sistema.

La mayor dificultad que presenta este sistema en una implementación real es el despliegue de una cadena de bloques con suficientes nodos. Para asegurar una distribución equilibrada de la potencia de cómputo y evitar el control de la red, es fundamental contar con nodos independientes. Para ello se pueden contemplar diversas opciones, como llegar a acuerdos con entidades tanto públicas como privadas, o con universidades y centros de investigación. En cualquier caso, es imprescindible garantizar una red heterogénea y resistente a intentos de corrupción.

El sistema electoral de una nación es una infraestructura crítica, que aunque actualmente resulte algo arcaico, sus problemas están bien estudiados y son pocas las vulnerabilidades que presenta. La implementación de un sistema en línea como el propuesto, requiere de un estudio exhaustivo para garantizar la seguridad del mismo. Sin embargo, este trabajo muestra su viabilidad, que con más investigación y pruebas, podría llegar a ser implementado.

Agradecimientos

Esta investigación ha sido posible gracias a las siguientes organizaciones y proyectos: SCITALA C064/23 y la Cátedra de Ciberseguridad ULL-INCIBE financiada a través del NextGeneration EU a través del Plan de Recuperación, Transformación y Resiliencia, el Ministerio de Universidades de España a través del Proyecto FIDEMOV, el proyecto PID2022-138933OB-I00, y el proyecto 2023DIG28 apoyado por CajaCanarias y la Fundación “la Caixa”.

Referencias

[1] S. Park, M. Specter, N. Narula y R. L. Rivest, “Going from bad to worse: from Internet voting to blockchain voting”, *Journal of Cybersecurity*, vol. 7, n.º 1, tyaa025, feb. de 2021, ISSN: 2057-2085. DOI: 10.1093/cybsec/tyaa025. eprint: https://academic.oup.com/cybersecurity/article-pdf/7/1/tyaa025/42533672/tyaa025.pdf.

[2] T. Elgamal, “A public key cryptosystem and a signature scheme based on discrete logarithms”, *IEEE Transactions on Information Theory*, vol. 31, n.º 4, págs. 469-472, 1985. DOI: 10.1109/TIT.1985.1057074.

[3] R. Zhou y Z. Lin, “An Improved Exponential ElGamal Encryption Scheme with Additive Homomorphism”, en *2022 International Conference on Blockchain Technology and Information Security (ICBCTIS)*, 2022, págs. 25-27. DOI: 10.1109/ICBCTIS55569.2022.00017.

[4] R. L. Rivest, A. Shamir y L. Adleman, “A method for obtaining digital signatures and public-key cryptosystems”, *Commun. ACM*, vol. 21, n.º 2, págs. 120-126, feb. de 1978, ISSN: 0001-0782. DOI: 10.1145/359340.359342.

[5] S. Goldwasser, S. Micali y C. Rackoff, "The Knowledge Complexity of Interactive Proof Systems", *SIAM Journal on Computing*, vol. 18, n.º 1, págs. 186-208, 1989. DOI: 10.1137/0218012. eprint: https://doi.org/10.1137/0218012. dirección: https://doi.org/10.1137/0218012.

[6] G. Maxwell y A. Poelstra, "Borromean Ring Signatures *", 2015. dirección: https://api.semanticscholar.org/CorpusID:16289588.

[7] A. Fiat y A. Shamir, "How to prove yourself: Practical solutions to identification and signature problems", en *Conference on the theory and application of cryptographic techniques*, Springer, 1986, págs. 186-194. DOI: 10.1007/3-540-47721-7_12.

[8] D. G. WOOD, *ETHEREUM: A SECURE DECENTRALISED GENERALISED TRANSACTION LEDGER SHANGHAI VERSION efc5f9a – 2025-02-04*, Github repository, 2025. dirección: https://ethereum.github.io/yellowpaper/paper.pdf.

[9] Y. Hao, "Research of the 51 % attack based on blockchain", en *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, 2022, págs. 278-283. DOI: 10.1109/CVIDLICCEA56201.2022.9824528.

[10] A. M. Don Johnson y S. Vanston, *The Elliptic Curve Digital Signature Algorithm (ECDSA)*, 2001. DOI: 10.1007/s102070100002.

[11] D. R. L. Brown, *SEC 2: Recommended Elliptic Curve Domain Parameter*, 2010. dirección: https://www.secg.org/sec2-v2.pdf.

[12] A. Shamir, "How to share a secret", *Commun. ACM*, vol. 22, n.º 11, págs. 612-613, nov. de 1979, ISSN: 0001-0782. DOI: 10.1145/359168.359176.

[13] A. Joux, A. Odlyzko y C. Pierrot, "The Past, Evolving Present, and Future of the Discrete Logarithm", en *Open Problems in Mathematics and Computational Science*, Ç. K. Koç, ed. Cham: Springer International Publishing, 2014, págs. 5-36, ISBN: 978-3-319-10683-0. DOI: 10.1007/978-3-319-10683-0_2.

[14] P. Shor, "Algorithms for quantum computation: discrete logarithms and factoring", en *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 1994, págs. 124-134. DOI: 10.1109/SFCS.1994.365700.

[15] V. Lyubashevsky, N. K. Nguyen y M. Plancon, *Lattice-Based Zero-Knowledge Proofs and Applications: Shorter, Simpler, and More General*, Cryptology ePrint Archive, Paper 2022/284, 2022. dirección: https://eprint.iacr.org/2022/284.

[16] L. K. Grover, "A fast quantum mechanical algorithm for database search", en *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, ép. STOC '96, Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 1996, págs. 212-219, ISBN: 0897917855. DOI: 10.1145/237814.237866.

[17] N. I. of Standards y Technology, *Module-Lattice-Based Key-Encapsulation Mechanism Standard*, 2024. DOI: 10.6028/NIST.FIPS.203. dirección: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.203.pdf.

[18] N. I. of Standards y Technology, *Module-Lattice-Based Digital Signature Standard*, 2024. DOI: 10.6028/NIST.FIPS.204. dirección: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.204.pdf.

[19] J. Fan y F. Vercauteren, *Somewhat Practical Fully Homomorphic Encryption*, Cryptology ePrint Archive, Paper 2012/144, 2012. dirección: https://eprint.iacr.org/2012/144.

[20] C. Gentry, C. Peikert y V. Vaikuntanathan, *Trapdoors for Hard Lattices and New Cryptographic Constructions*, Cryptology ePrint Archive, Paper 2007/432, 2007. dirección: https://eprint.iacr.org/2007/432.

Sesión Investigación VI: Ciberamenazas y detección avanzada

Generación de títulos en español para la alerta temprana de vulnerabilidades mediante Inteligencia Artificial Generativa

Javier Román Martínez*, David Triana Robles‡, Mouhcine El Oualidi Charchmi§, Ines Salamanca Estévez†, Noemí DeCastro-García¶
RIASC, Universidad de León
Campus de Vegazana s/n, 24071 León, España
* jromm@unileon.es, ‡ dtrir@unileon.es, § mouac@unileon.es, † isale@unileon.es, ¶ ncasg@unileon.es

***Resumen*—La gestión de la alerta temprana de vulnerabilidades desempeña un papel clave en el ámbito de la ciberseguridad, ya que permite minimizar el tiempo de exposición a posibles amenazas. Este estudio analiza el impacto de incluir sistemas de inteligencia artificial en la generación automatizada de títulos para alertas tempranas gestionadas por el Instituto Nacional de Ciberseguridad. Se evalúan diferentes herramientas de inteligencia artificial generativa, comparando la precisión, la coherencia semántica y la eficiencia de los títulos generados. A través del uso de modelos de representación vectorial de palabras y análisis estadístico, se identifican patrones en la estructura de los títulos y se determina qué herramienta ofrece los mejores resultados.**

***Index Terms*—Inteligencia Artificial Generativa, Ciberseguridad, Vulnerabilidades, Generación automática de contenido**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

La proliferación de vulnerabilidades en sistemas informáticos ha convertido la gestión de la alerta temprana en un elemento esencial para la ciberseguridad. Identificar y comunicar vulnerabilidades de manera eficiente permite a las organizaciones aplicar medidas preventivas rápidamente, minimizando el impacto de posibles ataques. Dado que los títulos son la primera referencia en la gestión de vulnerabilidades, una correcta estandarización de los mismos mejora la detección, análisis y respuesta temprana a incidentes de seguridad. En este contexto, la generación precisa y eficiente de títulos para reportes de vulnerabilidades es un factor clave para la correcta identificación y comprensión de los riesgos asociados.

El Instituto Nacional de Ciberseguridad (INCIBE) es responsable de la gestión de la alerta temprana en español de vulnerabilidades, un proceso que actualmente se lleva a cabo de manera manual. Este procedimiento comienza cuando el NIST (National Institute of Standards and Technology [1]) publica nuevas vulnerabilidades en la base de datos NVD (National Vulnerability Database [2]), proporcionando una descripción detallada en inglés. INCIBE accede a esta información y la incorpora a su propio *feed* de vulnerabilidades, el cual posteriormente es publicado en su web [3]. De forma paralela a esta integración, en la plataforma del NIST, INCIBE lleva a cabo la traducción manual de la descripción de la vulnerabilidad y la generación de un título adecuado en español. Una vez que esta fase ha concluido, la información traducida y revisada se actualiza en el *feed* interno y, finalmente, en la web de INCIBE, garantizando que los usuarios tengan acceso a una versión en español de las alertas de vulnerabilidades.

Este proceso, aunque funcional, resulta laborioso, lo que puede provocar demoras y variaciones en la calidad de los títulos generados. La aplicación de Inteligencia Artificial Generativa (IAG) en este contexto ofrece una oportunidad para optimizar la generación de títulos, garantizando mayor coherencia, reduciendo los tiempos de procesamiento y mejorando la eficiencia en la difusión de alertas de seguridad. Para que la aplicación de la IAG sea efectiva, es fundamental que el problema a abordar presente patrones de comportamiento que puedan ser identificados y aprendidos por los modelos [4]. En estos casos, los modelos pueden generalizar contenido a partir de ejemplos previos, optimizando su capacidad de respuesta y reduciendo la incertidumbre en sus predicciones. Al observar que los títulos siguen estructuras recurrentes, surge la idea de explorar la automatización del proceso mediante IAG.

En este artículo se presenta una evaluación preliminar de la eficiencia que presentan distintas herramientas de IAG en la generación de títulos de la alerta temprana en español de vulnerabilidades. Un aspecto fundamental en el proceso de generación de títulos es la traducción de las descripciones de las vulnerabilidades en las que se basan. Estas traducciones son esenciales para la construcción de los títulos. En estudios previos [5], se ha analizado la generación automática de traducciones y su impacto en la precisión y coherencia del contenido generado. De los resultados obtenidos, se ha extraído la selección de IAGs que se han analizado en este estudio: ChatGPT [6], Copilot [7] y ALIA [8]. La primera parte de la investigación se ha centrado en identificar patrones en la creación de títulos que permitieran obtener el *prompt* óptimo que sirva como *input* a cada IAG. Seguidamente, se ha pasado a generar los títulos de una muestra de vulnerabilidades para, finalmente, determinar la herramienta más adecuada para la tarea mediante análisis estadístico y de embebimiento de palabras, estableciendo criterios objetivos que faciliten su implementación en la gestión de alertas de vulnerabilidades.

Este artículo se organiza de la siguiente manera: la Sección II describe el diseño del experimento. En la Sección III, se presentan los hallazgos experimentales. Finalmente, se incluyen las conclusiones de este trabajo de investigación.

II. Sección experimental

En esta sección se describe el diseño experimental, incluyendo las preguntas de investigación, los datos utilizados, la selección de IAGs y la metodología para identificar patrones, generar títulos y evaluar su calidad mediante similitud

semántica.

II-A. Preguntas de investigación

El objetivo general de este trabajo es determinar el efecto de incluir IAG en el proceso de generación de títulos de la alerta temprana de vulnerabilidades que realiza INCIBE. Las preguntas de investigación que se quieren resolver en este trabajo se enumeran a continuación:

- PI1 ¿Existen patrones en los títulos de las descripciones de las vulnerabilidades? ¿Qué patrón es el más adecuado/-frecuente/útil?
- PI2 ¿Qué IAG aporta los títulos de mejor calidad?
- PI3 ¿Qué IAG es más eficiente en la tarea de generación de títulos para las descripciones de la alerta temprana de vulnerabilidades?

II-B. Conjuntos de datos

Se utilizó un *dataset* proporcionado por INCIBE [9], el cual contiene un total de 205282 vulnerabilidades, cada una con su respectiva descripción en inglés, su traducción al español y su título. Este conjunto de datos sirvió como base para la identificación de patrones en los títulos de las vulnerabilidades y el análisis de similitud semántica.

Debido a las limitaciones de las IAGs, que impedían la lectura de archivos demasiado grandes, se estudiaron específicamente las primeras 2269 vulnerabilidades con título para la identificación de patrones de forma automática.

Por otro lado, para llevar a cabo las pruebas de generación de títulos, se seleccionó una muestra compuesta por las primeras 22 vulnerabilidades con título registradas en la base de datos. Esta selección se realizó con el objetivo de garantizar que, dentro de las restricciones propias de las IAG utilizadas, el tiempo de procesamiento no resultara excesivo. Esto se debió a que, tanto por las limitaciones de las versiones gratuitas de las herramientas comparadas como por las del hardware disponible, la generación de respuestas por parte de cada IAG implicaba un consumo considerable de recursos y tiempo.

II-C. IAGs

En este estudio se emplean diversas IAGs con el objetivo de generar los títulos en español de las alertas tempranas de vulnerabilidades. La generación de títulos depende de forma inherente a la descripción de las vulnerabilidades. Por ese motivo, la selección se ha basado en otras investigaciones relacionadas [5]. Así, las traducciones han sido generadas por los sistemas de traducción automático e IAGs que se enumeran a continuación:

- ChatGPT [6] (Versión GPT-4o mini): Modelo de lenguaje avanzado desarrollado por OpenAI, capaz de realizar traducciones y generar texto en múltiples idiomas con alto nivel de fluidez y contexto.
- Bing Translator [10]: Servicio de traducción basado en IA de Microsoft, integrado en diversas plataformas y aplicaciones.
- DeepL [11]: Herramienta de traducción que emplea redes neuronales avanzadas para ofrecer traducciones de alta calidad con especial énfasis en matices lingüísticos.
- Google Translate [12]: Traductor automático de Google, ampliamente utilizado y mejorado mediante aprendizaje automático y redes neuronales.
- Copilot For Security (CFS) [13]: Versión especializada de Copilot enfocada en la seguridad informática, con capacidades avanzadas de análisis y generación de texto técnico.
- Copilot [7]: Asistente de IA desarrollado por Microsoft basado en modelos de OpenAI, diseñado para asistir en la programación y otras tareas relacionadas con el desarrollo de software.
- CFS1: CFS enriquecido con la base de datos de referencia [9]: Variante de CFS que trabaja con la base de datos utilizada como referencia, incluyendo las muestras analizadas, para mejorar la precisión en tareas específicas.
- CFS2: CFS enriquecido con la base de datos de referencia [9] sin las descripciones de la muestra.
- Salamandra [14]: Herramienta basada en IA utilizada en tareas de generación y traducción de contenido.

Excepto en CFS, CFS1 y CFS2, se han utilizado versiones gratuitas de todas las herramientas. Así, y basándonos en los resultados obtenidos en [5], las IAG seleccionadas para la generación de títulos han sido ChatGPT, Copilot y Salamandra.

Cabe destacar que también se ha llevado a cabo la experimentación partiendo de la descripción general en inglés.

II-D. Metodología

El flujo de trabajo incluyó:

1. Preprocesamiento de los datos de la muestra de INCIBE [9]: El procesamiento de los datos se llevó a cabo mediante un *script* en *Python* que extrajo todos los títulos existentes, eliminó duplicados y los guardó en un archivo de texto plano. Posteriormente, se realizó un análisis manual para identificar patrones en los títulos determinando inicialmente la existencia de seis estructuras distintas. Para realizar este análisis se tuvo en cuenta que los sistemas de alerta temprana suelen basarse en la recolección y análisis de informes de vulnerabilidades provenientes de diversas fuentes, como bases de datos de seguridad y reportes de investigadores (ej. National Vulnerability Database, NVD [2]; Common Vulnerabilities and Exposures, CVE [15]). Para facilitar su clasificación y priorización, los títulos de las vulnerabilidades deben ser claros y estructurados, incluyendo información clave como: Componente afectado (software, dispositivo, módulo), tipo de vulnerabilidad (XSS, SQL Injection, RCE, etc.), versión afectada (si aplica) y contexto del impacto (cómo y dónde se manifiesta la vulnerabilidad).
   Debido al gran número de títulos y el riesgo de error humano, se utilizó *ChatGPT 3.5* y *Copilot*, para ampliar el análisis y validar los grupos identificados manualmente. Esto permitió identificar patrones adicionales, refinando los grupos hasta obtener un total de nueve estructuras.
2. Generación de títulos mediante IAG: se solicitó a las IAGs generar un *prompt* óptimo para la creación de nuevos títulos basados en la descripción de una vulnerabilidad con una estructura que siguiera el patrón seleccionado. Con base en este *prompt*, se generaron

los títulos correspondientes de la muestra; es decir, se generaron títulos para 22 traducciones en español de vulnerabilidades. Este proceso se repitió con cada conjunto de traducciones creadas con los diferentes traductores automáticos explicados en la sección anterior y para la descripción original en inglés.
3. Cálculo de similitud semántica usando *word embeddings* [16].
4. Evaluación de los resultados con los títulos originales propuestos por INCIBE. Este enfoque permitió obtener una evaluación cuantitativa sobre la calidad de los títulos generados, asegurando una metodología robusta y reproducible para futuros estudios.

II-E. Análisis

Para evaluar la similitud semántica entre los títulos generados y los títulos originales, se utilizó la técnica de *word embeddings* junto con la similitud del coseno. Se entrenaron varios modelos de *word embeddings*, empleando diferentes corpus disponibles:

- Primer modelo: basado en todas las traducciones del conjunto de datos original de INCIBE [9].
- Segundo modelo: basado en un corpus de traducciones público proporcionado por INCIBE, el cual fue desarrollado en 2019 en colaboración con la National Vulnerability Database (NVD) del NIST con el objetivo de traducir al español las descripciones de vulnerabilidades recogidas en dicha base de datos. Este esfuerzo resultó en la creación de un corpus bilingüe con más de 100000 unidades de traducción, disponible públicamente en el repositorio ELRC-SHARE, y orientado a mejorar las capacidades de análisis y respuesta ante vulnerabilidades en el ámbito hispanohablante [17].
- Tercer modelo: utilizando modelos preentrenados (`wiki.en.bin` y `wiki.es.bin`) disponibles en *FastText* [18] y en el proyecto de investigación *MUSE* [19] desarrollado por Meta.

Para los dos primeros modelos, se creó un diccionario de *palabras ancla* extrayendo palabras comunes entre las traducciones en inglés y español. Este procedimiento se llevó a cabo mediante un *script* en *Python* que analizó y comparó los archivos de texto, identificando palabras coincidentes y registrando su frecuencia de aparición. Para el tercer modelo, se utilizó un diccionario de *palabras ancla* proporcionado por las fuentes de los modelos.

Cada modelo fue alineado mediante la técnica de Procrustes ortogonal [20], permitiendo calcular la similitud semántica entre las frases comparadas. La similitud se evaluó en un rango de [0,1], donde 0 indica la menor aproximación y 1 la mayor.

Una vez aplicados los *word embeddings*, se calcularon estadísticos descriptivos de los resultados obtenidos de toda la muestra.

II-F. Detalles Técnicos

El procesamiento y análisis de datos se realizaron con herramientas de *Python*, incluyendo las librerías:

- *FastText [18]*: Para el entrenamiento de modelos de *word embeddings*.
- *SciPy [21]*: Para el cálculo de la similitud del coseno.
- *NumPy* [22] y *Pandas* [23]: Para la manipulación y análisis de datos.

III. Resultados

III-A. PI1: Búsqueda de patrones en los títulos

Los patrones fueron analizados primero de forma manual, creando nuevas categorías en caso de que un título no encajara en ninguna de las existentes. Para el análisis con *Copilot* y *ChatGPT*, se utilizó el siguiente *prompt*, adjuntando un archivo JSON con los títulos a analizar:

> *Analiza cómo se han creado los títulos de estos textos, estudia si existen patrones.*

Los patrones encontrados, de forma manual, han sido:

1. Nombre de producto/software/dispositivo
   - Patrón: [Nombre del Producto/Software/Dispositivo]
   - Ejemplo: SAP CRM WebClient UI
2. Código de CVE
   - Patrón: [CVE]
   - Ejemplo: CVE-2018-20249
3. Descripción de vulnerabilidad
   - Patrón: vulnerabilidad de [tipo de vulnerabilidad] en [producto]
   - Ejemplo: Vulnerabilidad de inyección SQL en el módulo Sections para PHP-Nuke
4. Descripción de una función/archivo/módulo
   - Patrón: el/la [nombre de la función/archivo/módulo] en [contexto]
   - Ejemplo: el encabezado HTTP en los dispositivos VIVOTEK IP Camera
5. Componente específico en aplicaciones
   - Patrón: el/la [componente] en [aplicación]
   - Ejemplo: el formulario de inicio de sesión en Microweber CMS
6. Configuraciones y parámetros en aplicaciones/dispositivos
   - Patrón: el/los [parámetro(s)] en [aplicación/dispositivo]
   - Ejemplo: el parámetro system_time_timezone en la función SetNTPServerSettings en D-Link COVR 1200,1202,1203
7. Referencias a componentes y módulos en contextos específicos
   - Patrón: el/la [componente/módulo] en [contexto]
   - Ejemplo: el componente Outside In Technology en Oracle Application Server
8. Operaciones y acciones en productos/aplicaciones
   - Patrón: el/la [acción] en [producto/aplicación]
   - Ejemplo: la generación del código HTML relacionado con los emojis en GitLab CE/EE
9. Nombres específicos de funciones y métodos
   - Patrón: la/el [nombre de la función/método] en [contexto]
   - Ejemplo: la función ext4_xattr_check_entries en el kernel de Linux

A pesar de identificar estos patrones, no se ha encontrado un motivo claro para seleccionar un patrón específico u otro al generar el título basado en su descripción. Cabe destacar que a excepción de los 3 primeros grupos, los títulos siguen siempre la misma estructura:

El/La/Los/Las [nombre de elemento/s afectado/s] en [contexto en el que se ven afectados]

Aunque la fórmula "vulnerabilidad de [tipo de vulnerabilidad] en [producto]" proporciona suficiente información para considerarla una estructura de título válida, tiene una representación muy limitada.

Una vez determinado el patrón óptimo para la generación de títulos, se compartió este análisis con *Copilot* y *ChatGPT*, junto con el siguiente *prompt* para que cada uno generara su propia propuesta de *prompt* ideal:

> *Crea un prompt que pueda utilizarse para generar nuevos títulos a partir de nuevos textos.*

Las respuestas obtenidas se presentan en la Tabla I.

IAG	*Prompt* **propuesto**
ChatGPT	Genera un título conciso para un informe de vulnerabilidad basado en la siguiente descripción. El título debe comenzar con el nombre del archivo, software o sistema afectado, seguido de su ubicación o contexto si es relevante. No menciones la vulnerabilidad específica ni detalles técnicos complejos. Usa un formato similar a estos ejemplos: *'archivo_en_sistema'*, *'módulo en software'*, *'software en versión_afectada'*.
Copilot	Crea un título para un informe de vulnerabilidad siguiendo estos pasos: ▪ Identifica el componente o software afectado. ▪ Especifica el tipo de vulnerabilidad (e.g., XSS, SQL Injection, Buffer Overflow). ▪ Incluye la versión del software si es relevante. ▪ Utiliza preposiciones como *"en"* para indicar el contexto y *"para"* para indicar el propósito. ▪ Añade detalles adicionales entre paréntesis si es necesario.

Tabla I
Prompts PROPUESTOS PARA LA GENERACIÓN DE TÍTULOS DE VULNERABILIDADES

Tras la generación de los *prompts* óptimos por parte de *ChatGPT* y *Copilot*, se evaluó manualmente el desempeño de cada modelo en la generación de títulos, considerando criterios de precisión, claridad y alineación con la descripción de la vulnerabilidad. Cada modelo generó cinco títulos por *prompt*, y su adecuación fue analizada cualitativamente según estos criterios.

El *prompt* generado por *Copilot* fue seleccionado como el más adecuado, ya que sus respuestas presentaban una estructura más coherente con las instrucciones dadas. Además, cuando la descripción de la vulnerabilidad estaba en inglés, se especificó en el *prompt* que el título debía generarse en español para evitar generaciones incorrectas.

III-B. PI2: Evaluación de la calidad de los títulos

Los resultados obtenidos de realizar la similitud del coseno de los *word embedding* se pueden ver reflejados en las tablas II, III y IV

Con respecto al modelo 1, Copilot y Salamandra tienen un rendimiento superior, mientras que ChatGPT obtiene puntuaciones más bajas. Sin embargo, con el modelo 2, Salamandra

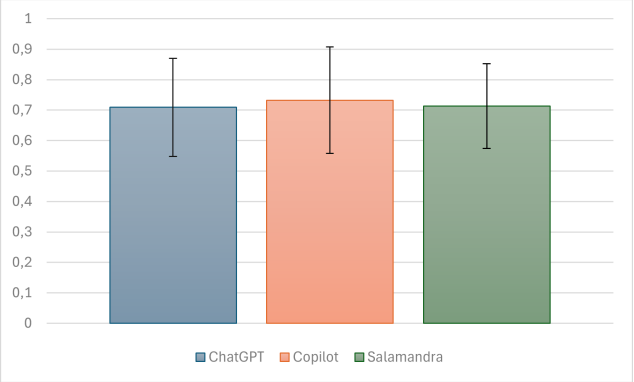


Figura 1. Media de los resultados obtenidos con los tres modelos, comparando los títulos generados a partir de la descripción traducida por INCIBE

lidera, seguido de Copilot, y ChatGPT se mantiene como el menos preciso. Por último, con el modelo 3, Copilot sigue siendo el mejor, con ChatGPT en último lugar. Las diferencias en los resultados pueden deberse a varios factores. Cada modelo de *word embeddings* representa el significado de las palabras de manera diferente, dependiendo de cómo fue entrenado. El modelo 1 está entrenado con vulnerabilidades traducidas a español por INCIBE, por lo que probablemente está más alineado con textos técnicos y de ciberseguridad. Por ese motivo, los títulos generados por modelos más especializados (como Copilot y Salamandra) pueden haberse ajustado mejor al contexto técnico, aportando mejores puntuaciones. Sin embargo, ChatGPT podría haber generado títulos más generales, lo que explicaría su menor rendimiento en este modelo. Por otra parte, el modelo 2 está entrenado con un corpus de ciberseguridad del año 2019, por lo que podría tener una representación semántica más amplia y menos actual. Con este modelo, Salamandra obtiene mejores resultados, lo que podría sugerir que su método de generación de títulos es más adaptable. ChatGPT sigue con puntuaciones más bajas. El modelo 3 está preentrenado con gran cantidad de textos diversos, sin un enfoque específico en ciberseguridad. En este escenario, Copilot sería el mejor generado, lo que indica que es un generador de títulos más robusto en contextos generales. ChatGPT y Salamandra tienen puntuaciones más bajas, lo que sugiere que sus estrategias de generación pueden no estar alineadas con el conocimiento general aprendido por este modelo.

De manera general, aunque los resultados puedan parecer similares (véase la Figura 1), se observan algunas diferencias en la media y la desviación típica obtenidas. No obstante, dado que estas diferencias no parecen significativas, se procede a analizar el coste de implementación en términos de tiempo. Este análisis permite identificar que Copilot presenta un rendimiento más estable y elevado en todos los modelos, lo que refleja su capacidad para extraer términos clave y estructurar los títulos con precisión. Además, se adapta bien tanto a modelos específicos como generales, lo que sugiere una buena consistencia semántica en su algoritmo. Por otro lado, Salamandra muestra un mejor desempeño en modelos específicos, lo que podría indicar que su generación de títulos depende más del contexto técnico. ChatGPT, en cambio, obtie-

Tabla II
RESULTADOS DE LA PRUEBA DE *word embeddings* USANDO EL MODELO ENTRENADO EN ESPAÑOL CON LAS VULNERABILIDADES TRADUCIDAS POR INCIBE.

Traductor	Generador de título	Media	Desv. típica	Mediana	Máximo	Mínimo	Tiempo medio (ms)
INCIBE	ChatGPT	0.7356	0.1552	0.7682	0.9161	0.3325	0.1322
	Copilot	0.7521	0.1709	0.8350	0.9833	0.3492	0.1343
	Salamandra	0.7444	0.1326	0.7796	0.9317	0.3809	0.1837
ChatGPT	ChatGPT	0.7236	0.1596	0.7392	0.9279	0.2871	0.1524
	Copilot	0.7430	0.1697	0.7707	0.9206	0.3018	0.1254
	Salamandra	0.7506	0.1479	0.7745	0.9480	0.3107	0.1474
Bing	ChatGPT	0.7162	0.1605	0.7377	0.9222	0.3001	0.1408
	Copilot	0.7267	0.1610	0.7798	0.9071	0.3172	0.1171
	Salamandra	0.7449	0.1297	0.7773	0.8939	0.3189	0.1534
DeepL	ChatGPT	0.7155	0.1575	0.7353	0.9211	0.3062	0.1323
	Copilot	0.7399	0.1694	0.7743	0.9282	0.3455	0.1205
	Salamandra	0.7502	0.1293	0.7866	0.8916	0.3029	0.1363
Google Translate	ChatGPT	0.7211	0.1537	0.7159	0.9265	0.3222	0.1317
	Copilot	0.7353	0.1668	0.7499	0.9724	0.3139	0.1200
	Salamandra	0.7253	0.1252	0.7500	0.8827	0.3699	0.1392
CFS	ChatGPT	0.7147	0.1596	0.7253	0.9115	0.3222	0.1266
	Copilot	0.7258	0.1668	0.7718	0.9100	0.3139	0.1142
	Salamandra	0.7562	0.1339	0.7821	0.9241	0.4432	0.1414
Copilot	ChatGPT	0.7102	0.1514	0.7000	0.9115	0.3222	0.1304
	Copilot	0.7432	0.1775	0.7990	0.9606	0.3022	0.1264
	Salamandra	0.7376	0.1611	0.7921	0.9106	0.2233	0.1329
CFS1	ChatGPT	0.7261	0.1520	0.7199	0.9273	0.3248	0.1281
	Copilot	0.7529	0.1740	0.8092	0.9760	0.3018	0.1309
	Salamandra	0.7310	0.1154	0.7359	0.8983	0.3557	0.1387
CFS2	ChatGPT	0.7334	0.1630	0.7595	0.9225	0.3008	0.1334
	Copilot	0.7423	0.1720	0.8141	0.9206	0.3018	0.1191
	Salamandra	0.7361	0.1308	0.7700	0.9010	0.4486	0.1509
Salamandra	ChatGPT	0.7720	0.1266	0.7952	0.9325	0.4026	0.1466
	Copilot	0.7864	0.1290	0.8054	0.9657	0.4386	0.1433
	Salamandra	0.7106	0.1322	0.7286	0.8823	0.2840	0.1485
Descripción original en inglés	ChatGPT	0.7899	0.1240	0.8298	0.9346	0.4038	0.1490
	Copilot	0.8107	0.1183	0.8399	0.9502	0.4136	0.1348
	Salamandra	0.7477	0.1148	0.7739	0.9172	0.3927	0.1964

ne el rendimiento más bajo en todas las pruebas, posiblemente debido a que genera títulos más creativos, lo que reduce su alineación semántica con la descripción de referencia.

Por otro lado, resulta evidente que, en general, la generación automatizada de títulos funciona mejor cuando se parte de la descripción original de las vulnerabilidades en inglés, véase Tabla V. Al partir de la traducción a español de la descripciones, se podría estar perdiendo información o alterando la estructura semántica de la oración. Aunque las diferencias no parecen significativas, la descripción original en inglés permite una mejor generación del título. Este resultado podría ser debido a que tanto ChatGPT como Copilot han sido entrenadas con una mayor cantidad de datos en inglés que en español, comprendiendo y procesando mejor los textos en este idioma. Es por ello que la diferencia no es tan acusada cuando utilizamos Salamandra.

Por otra parte, el análisis mediante *word embeddings* podría estar ocasionando sesgos debido a que estos modelos de procesamiento de lenguaje natural están mejor calibrados para detectar relaciones semánticas en inglés que en español.

III-C. PI3: Eficiencia y limitaciones

Salamandra-7b-instruct tiene un tiempo de generación de título por cada descripción medio de 43.33 segundos mientras que Copilot genera las traducciones en un tiempo medio de 0.91 segundos y ChatGPT en 1.42 segundos.

Al realizar las pruebas de generación de títulos con Salamandra-7b-instruct, se ha observado un peor rendimiento. A pesar de que el *prompt* instrucciones detalladas, Salamandra-7b-instruct no genera los títulos siguiendo dichas indicaciones. En numerosos casos, es necesario postprocesar la respuesta, ya que reescribe la descripción de la vulnerabilidad en lugar de crear un título adecuado, o bien produce títulos genéricos que requieren ser completados manualmente. Esto obliga a repetir el proceso de generación para obtener resultados utilizables. Estos problemas no se han observado en ChatGPT o Copilot.

IV. CONCLUSIONES

El uso de una asistente basado en IAG puede mejorar la eficiencia en el proceso de generación automática de títulos en español de la alerta temprana de vulnerabilidades.

Como conclusiones finales, Copilot es el mejor generador de títulos en términos de precisión. Salamandra ofrece resultados sólidos y consistentes y ChatGPT genera títulos menos alineados. Además, en base a los resultados obtenidos, se recomendaría realizar este proceso de manera previa a realizar la traducción a español.

Como trabajo futuro, sería relevante ampliar la muestra de estudio, tanto en cantidad como en el intervalo de fechas considerado, con el fin de comprobar si los ajustes se degradan a medida que se incrementa el volumen de contenido y analizar

Tabla III
RESULTADOS DE LA PRUEBA DE *word embeddings* USANDO EL MODELO ENTRENADO EN ESPAÑOL CON EL CORPUS DE 2019.

Traductor	Generador	Media	Desv. típica	Mediana	Máximo	Mínimo	Tiempo Medio
INCIBE	ChatGPT	0.7405	0.1490	0.7788	0.9135	0.3968	0.1302
	Copilot	0.7566	0.1655	0.8257	0.9779	0.3857	0.1340
	Salamandra	0.7589	0.1198	0.7852	0.9348	0.4778	0.1915
ChatGPT	ChatGPT	0.7339	0.1503	0.7532	0.9177	0.3730	0.1502
	Copilot	0.7513	0.1651	0.7715	0.9119	0.3712	0.1243
	Salamandra	0.7665	0.1305	0.7982	0.9478	0.4083	0.1543
Bing	ChatGPT	0.7221	0.1499	0.7446	0.9201	0.3794	0.1394
	Copilot	0.7349	0.1537	0.7846	0.9115	0.3992	0.1171
	Salamandra	0.7600	0.1166	0.7884	0.8943	0.3948	0.1591
DeepL	ChatGPT	0.7247	0.1510	0.7521	0.9075	0.3815	0.1304
	Copilot	0.7458	0.1627	0.7799	0.9071	0.3825	0.1192
	Salamandra	0.7614	0.1114	0.7756	0.9068	0.4488	0.1422
Google Translate	ChatGPT	0.7282	0.1485	0.7324	0.9152	0.4076	0.1311
	Copilot	0.7428	0.1619	0.7491	0.9693	0.3898	0.1188
	Salamandra	0.7425	0.0999	0.7441	0.9032	0.4672	0.1432
CFS	ChatGPT	0.7221	0.1521	0.7260	0.8928	0.4076	0.1260
	Copilot	0.7334	0.1579	0.7780	0.9081	0.3898	0.1139
	Salamandra	0.7729	0.1173	0.7916	0.9161	0.5119	0.1480
Copilot	ChatGPT	0.7186	0.1474	0.7261	0.8928	0.4076	0.1295
	Copilot	0.7593	0.1708	0.8061	0.9620	0.3747	0.1249
	Salamandra	0.7617	0.1236	0.7982	0.9082	0.4184	0.1367
CFS1	ChatGPT	0.7326	0.1447	0.7249	0.9173	0.4041	0.1266
	Copilot	0.7580	0.1665	0.7876	0.9707	0.3712	0.1290
	Salamandra	0.7455	0.1040	0.7616	0.9047	0.4223	0.1478
CFS2	ChatGPT	0.7440	0.1516	0.7859	0.9152	0.3853	0.1317
	Copilot	0.7448	0.1639	0.7949	0.9044	0.3712	0.1177
	Salamandra	0.7526	0.1071	0.7760	0.8907	0.5280	0.1540
Salamandra	ChatGPT	0.7899	0.1205	0.8215	0.9388	0.4936	0.1449
	Copilot	0.8036	0.1105	0.8149	0.9644	0.5307	0.1421
	Salamandra	0.7340	0.1096	0.7662	0.8864	0.4307	0.1573
Descripción original en inglés	ChatGPT	0.8029	0.1111	0.8303	0.9392	0.4752	0.1536
	Copilot	0.8288	0.1105	0.8546	0.9496	0.4606	0.1407
	Salamandra	0.7637	0.1043	0.7947	0.8972	0.4573	0.2039

cómo se comporta la generación de títulos en vulnerabilidades más recientes.

AGRADECIMIENTOS

Esta investigación resulta del Proyecto Estratégico *Ciencia de datos para un modelo de inteligencia artificial en ciberseguridad* (C073/23), fruto del convenio de colaboración suscrito entre el Instituto Nacional de Ciberseguridad (INCIBE) y la Universidad de León. Esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation).

REFERENCIAS

[1] NIST, "NIST," https://www.nist.gov/, Último acceso: 15 de enero de 2025.

[2] ——, "National vulnerability database," 2005, Último acceso: 21 de enero de 2025. [Online]. Available: https://nvd.nist.gov/

[3] INCIBE, "Vulnerabilidades," https://www.incibe.es/incibe-cert/alerta-temprana/vulnerabilidades, Último acceso: 7 de marzo de 2024.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[5] Javier Román Martínez, David Triana Robles, Mouhcine El Oualidi Charchmi, Ines Salamanca Estévez, and Noemí DeCastro-García, "Generative artificial intelligence and machine translators in spanish translation of early vulnerability cybersecurity alerts," *Applied Sciences*, vol. 15, no. 8, 2025.

[6] OpenAI, "ChatGPT 3.5," https://openai.com/en/, Último acceso: 17 de febrero de 2025.

[7] Microsoft, "Copilot," https://www.microsoft.com/en-us/microsoft-copilot, Último acceso: 17 de febrero de 2025.

[8] Gobierno de España, "ALIA - Artificial Inteligence," https://www.alia.gob.es/, Último acceso: 17 de febrero de 2025.

[9] INCIBE, "BBDD de vulnerabilidades," Último acceso: 6 de febrero de 2025. [Online]. Available: https://drive.google.com/file/d/1tJOP8R98DcEMCNh6whXkqTsXGFpoTYU2/view?usp=sharing

[10] Microsoft, "Bing translator," https://www.bing.com/translator?to=en&setlang=en, Último acceso: 17 de febrero de 2025.

[11] D. GmbH, "Deepl," https://www.deepl.com/en/translator, Último acceso: 17 de febrero de 2025.

[12] Google, "Google translate," https://translate.google.es/?hl=es&sl=fr&tl=es&op=translate, Último acceso: 5 de febrero de 2025.

[13] Microsoft, "Copilot for security," https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot, Último acceso: 17 de febrero de 2025.

[14] Aitor Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, José Javier Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, and M. Villegas, "Salamandra technical report," 2025, Último acceso: 17 de febrero de 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.08489

[15] MITRE Corporation, "Common Vulnerabilities and Exposures (CVE) - Introduction," 2021, Último acceso: 5 de febrero de 2025. [Online]. Available: https://cve.mitre.org/docs/cve-intro-handout.pdf

[16] P. Gupta, S. Shekhawat, and K. Kumar, "Unsupervised quality estimation without reference corpus for subtitle machine translation using word embeddings," in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 32–38, Último acceso: 18 de diciembre de 2024.

[17] INCIBE, "NVD Database Vulnerability Descriptions," 2019, Último acceso: 26 de noviembre de 2024. [Online]. Available: https://elrc-share.eu/repository/browse/descripciones-de-vulnerabilidades-de-la-bbdd-nvd/9dd01c02cfcc11e9913100155d026706700705608 07f458f8e50bd5273778d7a/

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 07 2016.

[19] F. Research, "Muse," https://github.com/facebookresearch/MUSE, Último acceso: 30 de septiembre de 2024.

Tabla IV
RESULTADOS DE LA PRUEBA DE *word embeddings* USANDO EL MODELO EN ESPAÑOL PREENTRENADO.

Traductor	Generador	Media	Desv. típica	Mediana	Máximo	Mínimo	Tiempo Medio
INCIBE	ChatGPT	0.6512	0.1789	0.6495	0.8727	0.2476	0.1923
	Copilot	0.6884	0.1893	0.7061	0.9176	0.2843	0.1888
	Salamandra	0.6356	0.1642	0.7000	0.8679	0.2722	0.2768
ChatGPT	ChatGPT	0.6669	0.1816	0.6847	0.9199	0.2623	0.2168
	Copilot	0.6935	0.1805	0.7111	0.9161	0.2618	0.1709
	Salamandra	0.6741	0.1516	0.7025	0.8873	0.3153	0.2181
Bing	ChatGPT	0.6449	0.1706	0.6531	0.8761	0.2648	0.1937
	Copilot	0.6533	0.1720	0.6438	0.8735	0.2720	0.1584
	Salamandra	0.6549	0.1720	0.7248	0.8577	0.2486	0.2248
DeepL	ChatGPT	0.6492	0.1764	0.6606	0.9062	0.2318	0.1783
	Copilot	0.6882	0.1904	0.7312	0.8907	0.2204	0.1621
	Salamandra	0.6563	0.1567	0.7064	0.9138	0.2632	0.1989
Google Translate	ChatGPT	0.6373	0.1771	0.6621	0.9130	0.2231	0.1786
	Copilot	0.6588	0.1875	0.6558	0.9201	0.2213	0.1593
	Salamandra	0.6378	0.1694	0.6925	0.8684	0.2798	0.1995
CFS	ChatGPT	0.6543	0.1706	0.6662	0.8723	0.2595	0.1706
	Copilot	0.6680	0.1742	0.6823	0.8918	0.2641	0.1526
	Salamandra	0.6638	0.1843	0.6751	0.8838	0.2307	0.2048
Copilot	ChatGPT	0.6395	0.1767	0.6613	0.8723	0.2450	0.1785
	Copilot	0.6919	0.1877	0.7179	0.9154	0.2576	0.1691
	Salamandra	0.6886	0.1474	0.7377	0.8700	0.2799	0.1920
CFS1	ChatGPT	0.6421	0.1731	0.6456	0.8961	0.2642	0.1728
	Copilot	0.7018	0.1893	0.7169	0.9391	0.2618	0.1780
	Salamandra	0.6351	0.1451	0.6723	0.8402	0.2499	0.1985
CFS2	ChatGPT	0.6738	0.1736	0.6844	0.8996	0.2677	0.1809
	Copilot	0.6807	0.1842	0.7204	0.9135	0.2618	0.1585
	Salamandra	0.6638	0.1642	0.7166	0.8784	0.2439	0.2214
Salamandra	ChatGPT	0.6787	0.1513	0.7188	0.8725	0.2989	0.2111
	Copilot	0.7084	0.1355	0.7261	0.8881	0.3417	0.2015
	Salamandra	0.6233	0.1526	0.6748	0.8434	0.2615	0.2168
Descripción original en inglés	ChatGPT	0.7191	0.1428	0.7548	0.8919	0.3114	0.2259
	Copilot	0.7705	0.1207	0.7871	0.9123	0.3715	0.1981
	Salamandra	0.6676	0.1312	0.6841	0.8918	0.3046	0.3018

Tabla V
COMPARACIÓN DE LAS DIFERENCIAS PROMEDIO ENTRE LA VALORACIÓN OBTENIDA POR LOS TÍTULOS EN FUNCIÓN DE LA DESCRIPCIÓN UTILIZADA.

Modelo	Generador	Descripción en inglés	Descripciones en español (Promedio)	Diferencia
1	ChatGPT	0.7899	0.72684	0.06306
	Copilot	0.8107	0.74476	0.06594
	Salamandra	0.7477	0.73869	0.00901
2	ChatGPT	0.8029	0.73566	0.06724
	Copilot	0.8288	0.75305	0.07575
	Salamandra	0.7637	0.7556	0.0081
3	ChatGPT	0.7191	0.65379	0.06531
	Copilot	0.7705	0.6833	0.0872
	Salamandra	0.6676	0.65333	0.01427

[20] P. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.
[21] S. Developers, "Scipy: Open source scientific tools for Python," Último acceso: 26 de noviembre de 2024. [Online]. Available: https://www.scipy.org/
[22] N. Developers, "Numpy: The fundamental package for scientific computing with Python," Último acceso: 3 de marzo de 2025. [Online]. Available: https://numpy.org/
[23] P. Developers, "Pandas: Python Data Analysis Library," Último acceso: 3 de marzo de 2025. [Online]. Available: https://pandas.pydata.org/

ARGOS: Herramienta para la simulación de rutas de ataque en entornos de conciencia cibersituacional

Carmen Sánchez-Zas, Samuel García Sánchez, Xavier Larriva-Novo, Sonia Solera-Cotanilla, Oscar Jover-Walsh, Víctor A. Villagrá
Universidad Politécnica de Madrid (UPM). DIT, ETSI Telecomunicaciones. Avda. Complutense 30, 28040 Madrid
carmen.szas@upm.es, samuel.garcia.sanchez@alumnos.upm.es, xavier.larriva.novo@upm.es, sonia.solera@upm.es, oscar.jwalsh@upm.es, victor.villagra@upm.es

***Resumen*—En un contexto donde las amenazas de ciberseguridad son cada vez más sofisticadas, este artículo propone una solución para la prevención de ciberataques mediante la visualización del camino de ataque. Esto facilita la identificación de vulnerabilidades y la evaluación del impacto potencial de una intrusión, permitiendo a los equipos de seguridad anticiparse a posibles amenazas y fortalecer la protección de los sistemas. En esta propuesta se simula una ruta de ataque partiendo de técnicas definidas en MITRE ATT&CK y, a continuación, se representa gráficamente el posible recorrido que un adversario seguiría para comprometer un activo vulnerable desde cualquier punto de entrada al sistema. Al obtener acceso inicial a una infraestructura organizacional, un atacante puede moverse lateralmente entre activos hasta alcanzar los elementos más críticos. Para tratar de evitarlo se ha desarrollado una herramienta denominada ARGOS, presentada como un enfoque proactivo para mejorar la ciberseguridad en entornos de conciencia cibersituacional.**

***Index Terms*—Ruta de ataque, Ciberseguridad, TTPs, Cadenas de Markov, MITRE ATT&CK.**

Tipo de contribución: *Investigación original (límite 8 páginas)*

I. Introducción

Desde hace tiempo, las organizaciones y los gobiernos se enfrentan constantemente a posibles ataques de seguridad [1]. Sin embargo, la necesidad de una ciberdefensa de próxima generación se ha vuelto aún más urgente en esta era en la que las superficies de ataque que los ciberdelincuentes pueden explotar han crecido a un ritmo alarmante. Este rápido crecimiento deriva del aumento de la complejidad en los sistemas empresariales y la adopción masiva de servicios en la nube y dispositivos conectados [2]. En este contexto, la ciberdefensa basada en el análisis predictivo resulta más proactiva que las tecnologías actuales, que dependen de la detección de intrusiones.

Este nuevo enfoque de ciberdefensa está transformando la manera en que las organizaciones comprenden sus vulnerabilidades y se preparan ante posibles ataques. El *Informe de Situación Ciberseguridad en España 2024* [3] destaca que la adopción de la inteligencia artificial se ha acelerado en el ámbito de la ciberseguridad, proporcionando herramientas de análisis predictivo y detección automatizada de amenazas dentro de los entornos de conciencia cibersituacional. Ya no se trata solo de identificar vulnerabilidades aisladas o puntos débiles, ahora se trata de entender cómo los atacantes pueden encadenar estos puntos débiles para crear un camino de ataque crítico hacia sus activos más valiosos, como los administradores de dominio, bases de datos críticas, controladores de dominio de *Active Directory*, etc.

De este contexto surge esta propuesta, con el que se pretende ofrecer una solución para la prevención de ciberataques mediante la visualización del camino de ataque o *attack path visualization*, que es una representación gráfica de los posibles caminos de ataque que un adversario podría seguir para comprometer un activo desde cualquier punto de entrada en el sistema objetivo. Por ejemplo, una vez que un atacante ha logrado un acceso inicial al entorno organizacional, el camino de ataque muestra la posible ruta que el adversario puede tomar entre los activos para llegar a los elementos más valiosos, como los controladores de dominio.

El enfoque de la propuesta permite identificar los activos más vulnerables y evaluar la severidad de los ataques según rutas de ataque obtenidas de la simulación e integrar estos resultados con estrategias de mitigación, proporcionando recomendaciones adaptadas a los ataques identificados. De esta forma, se ha desarrollado un prototipo de esta propuesta en la herramienta ARGOS que contribuye a la mejora en la toma de decisiones en relación con la gestión de riesgos, un concepto clave en los entornos de conciencia cibersituacional.

A lo largo del documento presentaremos el entorno de la propuesta y los conceptos principales en los que se basa el desarrollo. A continuación, detallaremos la arquitectura de la herramienta, el desarrollo y los resultados obtenidos. En la sección I se presenta el entorno del trabajo. A continuación, en la sección II, se presentan trabajos anteriores que tratan la misma problemática. En la sección III se definen brevemente los conceptos teóricos relacionados con el proyecto. La arquitectura de ARGOS se define en la sección IV y se desarrolla en la sección V. Finalmente, la validación y los resultados obtenidos se presentan en la sección VI. Finalmente, las conclusiones y líneas futuras se presentan en la sección VII.

II. Trabajos Relacionados

La simulación de rutas de ataque en ciberseguridad mediante modelos probabilísticos ha sido un objetivo recurrente en la investigación de seguridad. Ha sido ampliamente estudiado para mejorar la gestión de riesgos y la identificación de activos críticos. En particular, el uso de cadenas de Markov ha demostrado ser una herramienta eficaz para estimar rutas de ataque y apoyar a la toma de decisiones en seguridad informática. En [4], los autores plantean el uso de HMM siguiendo una mezcla de distribuciones gaussianas (*Gaussian Mixture Hidden Markov Model* o GM-HMM) para conseguir detectar ataques multipaso, es decir, una secuencia de ataques.

Holgado et al. presentan en [5] un sistema de predicción de ataques multietapa basado en modelos ocultos de Markov

(HMM). Este enfoque modela la progresión de ataques partiendo de alertas que provienen de sistemas de detección de intrusiones (IDS), con el objetivo de predecir las próximas acciones de los atacantes y mejorar las estrategias de defensa. Sin embargo, se centra en la predicción, sobre todo centrada en ataques DDoS, pero no se integra en una herramienta enfocada en la gestión de riesgos y la protección específica de activos.

Siguiendo con la aplicación de HMM en este ámbito, Amin et al. combinan en [6] esta herramienta con técnicas de ciberengaño para frustrar el movimiento lateral de atacantes dentro de la red. Con un enfoque a la vez reactivo y proactivo, en primer lugar predicen la ruta de ataque más probable partiendo de alertas de IDS y determinan el despliegue óptimo de nodos señuelo a lo largo de esa ruta. Mediante la planificación parcialmente observable de Monte-Carlo (POMCP) definen acciones defensivas para bloquear al atacante en caso de que no siga la ruta calculada. Su investigación, partiendo de una base muy similar a la que se va a presentar en este documento, propone un enfoque hacia la defensa activa en lugar de la evaluación del impacto de los ataques en activos y la gestión de riesgos. Además, en [7] los autores tratan de predecir ataques y la familia a la que pertenecen mediante HMM. El objetivo es generar una secuencia de estados que correspondan con las etapas del ataque *Action Spoofing*, aunque sería generalizable a cualquier tipo de ataque.

Los autores de [8], por su parte, proponen un sistema basado en grafos de ataque y cadenas de Markov para modelar escenarios de ataque. Su metodología permite identificar los posibles caminos entre un origen y un destino dentro de una red y calcula la probabilidad de ataque. Finalmente, calcula el riesgo total del ataque, lo que le permite evaluar el nivel de peligro frente al grado de penetración del ataque. En este caso, los activos más vulnerables no se identifican y por tanto no se llega a evaluar la efectividad de diferentes estrategias de mitigación.

Otro enfoque en el cálculo de patrones de ataque consiste en considerar la variable temporal, como en [9], donde los autores presentan una herramienta capaz de computar todas las rutas de ataque iterando sobre las posibles duraciones de los posibles ataques. Los autores plantean añadir sobre los árboles de ataque utilizados para calcular las rutas, sellos temporales mediante *Boolean logic Driven Markov Processess* (BDMP).

El estudio sobre los trabajos relacionados han permitido demostrar que las cadenas de Markov son una herramienta clave para determinar qué activos son críticos, evaluar la severidad de los ataques, definiendo las respuestas más adecuadas para mitigar de forma proactiva las amenazas.

Los trabajos anteriormente analizados han contribuido notablemente al desarrollo de modelos predictivos en seguridad, pero su integración con enfoques de gestión de riesgos sigue siendo un desafío. A diferencia de estos, la herramienta que se propone en esta investigación contiene un modelo para la simulación de rutas de ataque utilizando cadenas de Markov con el objetivo de estimar qué movimientos son los más probables dentro de la red y utilizar esta información como complemento a la gestión de riesgos, lo que supone una ventaja con respecto a otros enfoques tradicionales.

III. Marco Teórico

III-A. MITRE ATT&CK

MITRE ATT&CK (*Adversarial Tactics, Techniques, and Common Knowledge*) [10] es una base de conocimientos universalmente accesible y continuamente actualizada para modelar, detectar, prevenir y combatir amenazas de ciberseguridad basándose en el comportamiento conocido de los ciberdelincuentes adversarios. Creado por MITRE *Corporation*, este marco de referencia ofrece una visión estructurada de las tácticas y técnicas empleadas por los atacantes en diferentes fases de una intrusión. Su principal objetivo es proporcionar a organizaciones, analistas de seguridad y profesionales de IT (*Information Technology*) una herramienta para comprender, detectar y mitigar amenazas cibernéticas de manera más efectiva.

El marco se organiza en matrices, que clasifican tácticas y técnicas en función de su propósito y cómo se relacionan con las etapas de un ataque. Las tácticas representan las metas o propósitos generales que los atacantes buscan alcanzar durante cada fase del ataque, como obtener acceso inicial, moverse lateralmente por la red o exfiltrar datos. Por otro lado, las técnicas describen los métodos específicos que los atacantes emplean para lograr las tácticas, por ejemplo, el aprovechamiento de vulnerabilidades o la ejecución de scripts maliciosos.

III-B. Cadenas de Markov

Una cadena de Markov [11] es un modelo matemático utilizado para describir sistemas que evolucionan de manera secuencial, donde la probabilidad de transición entre estados depende únicamente del estado actual y no del historial previo. Este principio se conoce como la propiedad de Markov.

Las cadenas de Markov son particularmente útiles para modelar y analizar procesos estocásticos, es decir, aquellos que involucran incertidumbre. Su capacidad para representar dinámicas donde el futuro depende solo del presente las hace valiosas en áreas como la predicción, la identificación de patrones de comportamiento y la simulación de escenarios en distintos contextos.

Una cadena de Markov consta de tres elementos principales: los estados, que representan las posibles condiciones del sistema (por ejemplo, “soleado”, “lluvioso” o “nublado” en un modelo climático); las transiciones, que describen los cambios entre estados con ciertas probabilidades; y la matriz de transición, que indica la probabilidad de pasar de un estado a otro, donde cada fila representa un estado actual y cada columna un estado futuro. Un estado es absorbente si, una vez alcanzado, no se puede abandonar, mientras que el estado inicial es aquel en el que se encuentra el sistema al inicio del análisis.

III-C. Rutas de ataque

Una ruta de ataque [12] es una representación visual del camino que un atacante podría seguir para explotar debilidades en un sistema y comprometer activos críticos. Este concepto, clave en la ciberseguridad, no solamente permite analizar vulnerabilidades aisladas, sino también cómo estas se interconectan para formar un trayecto explotable.

Su objetivo principal es ayudar a las organizaciones a identificar, priorizar y mitigar riesgos, abordando las amenazas

desde la perspectiva del atacante. En lugar de centrarse únicamente en solucionar vulnerabilidades individuales, el análisis de rutas de ataque proporciona un enfoque estratégico para visualizar el contexto completo de los riesgos, lo que resulta fundamental para proteger activos esenciales.

III-D. Bases de datos orientadas a grafos

Una ruta de ataque pierde gran parte de su potencial si no se visualiza correctamente. La visualización de rutas de ataque [13] es una herramienta poderosa que, mediante el uso de grafos, permite mapear los caminos que un atacante podría seguir dentro de una red, identificando conexiones entre activos, configuraciones erróneas y vulnerabilidades.

Esta representación visual facilita la priorización de recursos, ya que permite identificar los trayectos que conducen a activos críticos, como controladores de dominio o bases de datos sensibles. Así, la información sobre riesgos se transforma en una estrategia clara de mitigación.

Para mapear estas rutas de manera efectiva, se utilizan bases de datos de grafos, que representan de forma estructurada y visual la interconexión entre activos, vulnerabilidades y posibles movimientos de un atacante dentro de una red. Estas bases de datos se componen de nodos o vértices, que representan elementos individuales del sistema, y aristas o relaciones, que reflejan las conexiones entre ellos.

A diferencia de las bases de datos tradicionales, que organizan la información en tablas y requieren consultas complejas para analizar relaciones entre datos, las bases de datos de grafos están diseñadas para manejar interconexiones de manera eficiente. Este aspecto las hace especialmente útiles para modelar escenarios de ataque, ya que permiten identificar caminos críticos con mayor rapidez y precisión.

IV. Arquitectura de ARGOS

El diseño de la propuesta para la simulación de rutas de ataque se muestra en la Figura 1, cuya arquitectura se implementa en la herramienta ARGOS.

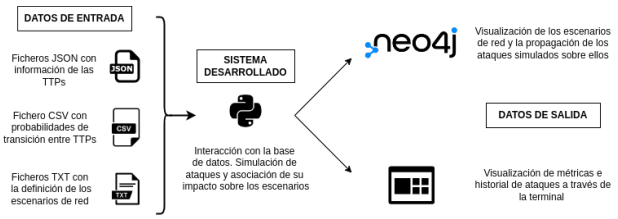


Figura 1. Diagrama de la arquitectura

La información sobre Tácticas, Técnicas y Procedimientos (TTPs) se extrae de la matriz empresarial de MITRE ATT&CK y se gestiona mediante archivos en formatos `.json`, y `.csv`, lo que facilita la actualización y la adaptación a nuevas amenazas sin necesidad de modificar la estructura del sistema.

Los escenarios de red se definen y cargan desde ficheros de texto que contienen las relaciones y configuraciones de los activos en un formato interpretable por Neo4j. Esto permite modelar la infraestructura de la red de forma escalable, pudiendo adaptarse a diferentes situaciones. Para la validación del sistema, se ha desarrollado un breve catálogo de tres simples escenarios: una pequeña oficina corporativa, un escenario de tipo *smart home* y una pequeña planta industrial.

Con toda esta información, el sistema simula rutas de ataque utilizando cadenas de Markov, generando secuencias de técnicas en función de probabilidades definidas. Posteriormente, estas secuencias se integran en el grafo del escenario de red que se haya cargado, permitiendo visualizar la propagación de amenazas y su impacto en los activos comprometidos. Esta asociación se lleva a cabo mediante un mapeo entre las TTPs y los activos afectados, considerando factores como la plataforma y capacidades del activo, los permisos de los que dispone y vulnerabilidades asociadas (CWE/CVE).

Para la visualización en la terminal, se utiliza la librería Rich, que proporciona un dashboard estructurado donde se muestra un resumen del ataque junto a métricas clave acerca de los activos y técnicas involucrados en el mismo, en base a las cuales se calcula una criticidad. Además, el sistema cuenta con un historial de ataques simulados.

V. Desarrollo

V-A. Modelo de datos

Como ya se ha mencionado, las bases de datos orientadas a grafos permiten representar la información mediante vértices o nodos y aristas o relaciones, lo que resulta especialmente útil para modelar redes informáticas y sus relaciones. Por ello, se ha desarrollado un modelo de datos escalable y flexible, permitiendo la representación tanto de distintos escenarios de red como secuencias de ataque bajo una lógica común.

Para el modelado de los escenarios de red, se han definido dos entidades principales: `Usuario` y `Activo`. La entidad `Usuario` representa las cuentas personales o administrativas que requieren autenticación en distintos elementos de la red. Por otro lado, la entidad `Activo` agrupa todos los dispositivos y sistemas de la infraestructura, dividiéndose en cuatro subcategorías según su función. Un `Servidor` representa dispositivos que ejecutan servicios accesibles, mientras que un `DispositivoRed` incluye equipos de usuario y otros sistemas conectados. Los elementos de seguridad, como *firewalls*, VPNs o sistemas IDS, se agrupan bajo la categoría `DispositivoSeguridad`. Finalmente, el software instalado en servidores o estaciones de trabajo se modela como `Aplicación`.

Cada nodo en el grafo cuenta con atributos en formato clave-valor, lo que permite almacenar información detallada sobre sus características. Estos atributos son fundamentales para asociar los activos con posibles vulnerabilidades y técnicas de ataque, facilitando el análisis de seguridad.

Las relaciones entre nodos son esenciales para modelar la conectividad y las dependencias dentro de la red. Se han definido cuatro tipos principales de relaciones entre los elementos presentes en los escenarios de red: `CONEXIÓN`, que indica la comunicación entre dos activos a través de un puerto; `PROTECCIÓN`, que señala cuando un activo está resguardado por un dispositivo de seguridad; `AUTENTICACIÓN`, que representa la necesidad de un usuario de autenticarse en un activo; y `ALOJAMIENTO`, que vincula una aplicación o servicio con el servidor donde está instalado.

Por otro lado, para modelar los ataques dentro del sistema, se ha definido la entidad `Técnica`, que representa cada una

de las técnicas de la matriz empresarial de MITRE ATT&CK utilizadas en una secuencia de ataque. Esta entidad se asocia con los activos vulnerables a través de atributos específicos. Además, se han definido dos relaciones clave para representar la progresión de los ataques: `TRANSICIÓN`, que indica la sucesión entre técnicas dentro de una secuencia de ataque, y `EXPLOTACIÓN`, que muestra la afectación de un activo por una técnica específica. En la Figura 2 se muestran todos los nodos y relaciones descritos.

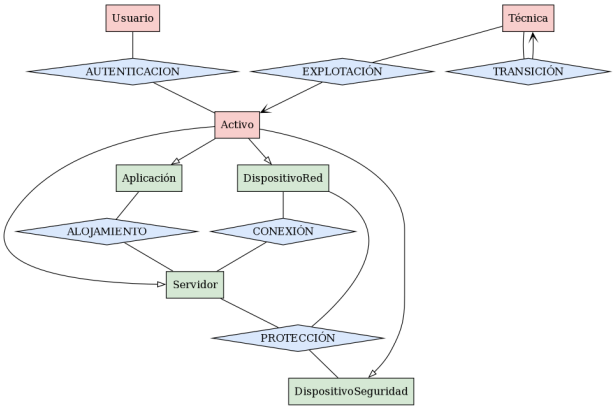


Figura 2. Modelo de datos empleado en ARGOS

V-B. Funcionamiento general

ARGOS es una herramienta de línea de comandos desarrollada en Python que permite analizar cómo una serie de TTPs pueden afectar un entorno de red simulado. Su principal objetivo es proporcionar una manera estructurada y automatizada de modelar ciberataques y visualizar su impacto en infraestructuras digitales.

ARGOS trabaja sobre una base de datos en Neo4j, que ha de ser configurada previamente. Para operar, ARGOS cuenta con cuatro comandos principales. El comando `prepare` permite cargar un escenario de red en la base de datos, transformándolo en un grafo dentro de Neo4j que representa activos y sus conexiones. Los escenarios se leen desde un archivo de texto ubicado en la carpeta `scenarios`. Estos archivos están escritos en un lenguaje denominado *Cypher*.

Cypher es el lenguaje de consulta diseñado específicamente para trabajar con grafos en Neo4j. Utiliza una sintaxis inspirada en SQL pero optimizada para manejar nodos y relaciones. De esta forma, es posible la realización de consultas y operaciones complejas sobre los grafos.

Cada escenario contiene información sobre los activos que lo componen (servidores, estaciones de trabajo, dispositivos, etc.) y sus características clave, como los sistemas operativos que utilizan, los permisos que poseen y las vulnerabilidades que podrían explotarse, utilizando el modelo de datos descrito en la sección V-A.

Antes de ejecutar cualquier simulación, ARGOS requiere que se cargue un escenario de red en la base de datos. Luego, el comando `attack` genera una secuencia de ataque basada en un modelo de Cadenas de Markov, que permite generar secuencias de técnicas de manera probabilística, y la ejecuta sobre el escenario cargado. Partiendo de una técnica seleccionada de manera aleatoria, la Cadena de Markov implementada permite calcular el siguiente estado únicamente a partir del estado actual utilizando las probabilidades de transición entre estados, definidas en el fichero `transitions.csv`. Esto significa que cada técnica utilizada en el ataque tiene una probabilidad de llevar a la siguiente, asegurando que las transiciones entre estados sean coherentes con patrones de ataque reales. De esta forma, ARGOS puede modelar la progresión de un ciberataque desde su punto de inicio hasta su posible finalización, considerando factores como la probabilidad de éxito de cada técnica y las condiciones del escenario.

Una vez generada la secuencia de ataque, ARGOS asocia cada técnica con los activos a los que afecta dentro del escenario cargado. Para ello, analiza características como los sistemas operativos presentes, los permisos disponibles y las vulnerabilidades conocidas, que habían sido definidas como atributos en los nodos. Si un activo cumple con los criterios necesarios, se establece una relación entre la técnica de ataque y el activo comprometido, reflejando el impacto del ataque en la red simulada. En la Figura 3, se presentan los atributos que caracterizan cada entidad, así como las condiciones para que se produzca una relación de tipo `EXPLOTACIÓN` entre ellos.

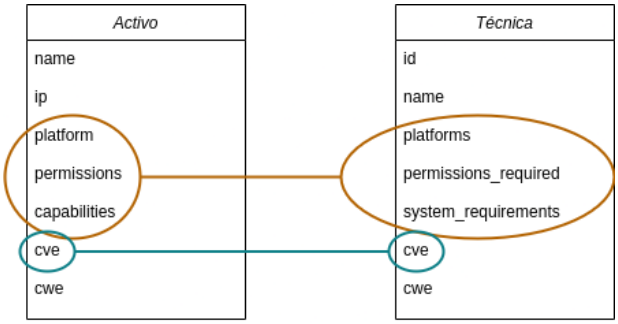


Figura 3. Condiciones de explotación de una técnica sobre un activo

La lógica utilizada para vincular las técnicas de ataque con los activos afectados se basa en la ejecución de una consulta Cypher sobre la base de datos Neo4j. Esta asociación se puede dar mediante dos condiciones. El filtro principal, representado de color naranja en la Figura 3, exige que el activo se ejecute en alguna de las plataformas afectadas por la técnica y, además, que cumpla con dos condiciones simultáneas: que posea al menos un permiso que coincida con los permisos requeridos por la técnica y que cuente con alguna capacidad del sistema que corresponda con los requisitos de la técnica. Esta condición compuesta refleja la necesidad de que el activo no solo opere en la plataforma objetivo, sino que también cumpla con las condiciones de acceso y recursos que la técnica requiere para ser aplicada. Adicionalmente, como alternativa a este conjunto de condiciones, la consulta también considera un segundo filtro (representado en color azul) para los activos que tengan vulnerabilidades identificadas por los CVEs relacionados con la técnica, permitiendo así que la técnica explote directamente vulnerabilidades conocidas.

Finalmente, el comando `history` permite consultar los ataques realizados previamente, mientras que el comando `clean` limpia la base de datos, eliminando cualquier información previa para realizar nuevas simulaciones desde cero.

VI. Validación

En esta sección, se presenta la validación de la herramienta ARGOS y los resultados obtenidos a partir de su ejecución. El proceso de validación se divide en varios pasos, que incluyen la carga de escenarios, la ejecución de simulaciones de ataque y el análisis de los resultados obtenidos, tal y como se muestra en el diagrama de ejecución de la Figura 4.

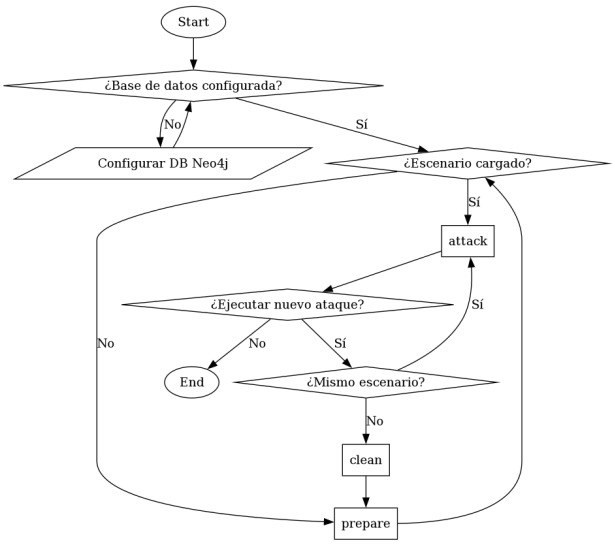


Figura 4. Diagrama de flujo de ARGOS

El primer paso antes de utilizar ARGOS es configurar e iniciar la base de datos. Una vez configurada, es necesario abrir la consola de Neo4j, ya que es donde se mostrarán los grafos correspondientes a la ejecución de ARGOS. Una vez hecho esto, ya se puede interactuar con la base de datos desde la terminal, mediante la interfaz de línea de comandos. El comando a ejecutar para cargar el escenario correspondiente a la oficina corporativa sería el siguiente: `$ python3 argos.py prepare scenarios/oficina.cypher`. Tras ejecutarlo, el programa indica que dicho escenario se ha cargado sin pro- blemas en la base de datos.

Tras este mensaje, en la consola de Neo4j se puede visualizar el escenario cargado en forma de grafo, tal y como se muestra en la Figura 5.

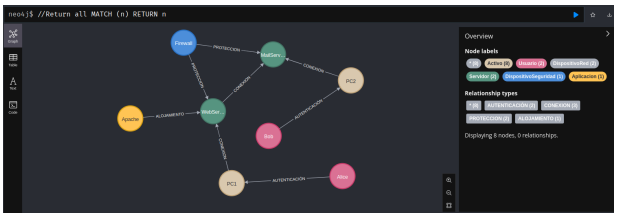


Figura 5. Escenario de oficina corporativa cargado en Neo4j

En este escenario, los usuarios están representados por los nodos `Alice` y `Bob`. `Alice` utiliza un PC Windows con permisos estándar y diversas capacidades que le permiten interactuar con otros activos en la red. El usuario `Bob` cuenta con Linux en su equipo, donde dispone de permisos de administrador. Tiene instalado el software *Digitaldesign CMS*. El `CVE-2009-3597` afecta a *Digitaldesign CMS* en su versión v0.1, que guarda información sensible en el directorio web raíz con insuficiente control de acceso, lo que permite a los atacantes remotos descargar el fichero de la base de datos a través de una petición directa a `autoconfig.dd`.

Los nodos de tipo servidor son `WebServer` y `MailServer`. `WebServer` corre en Windows con privilegios administrativos y capacidades relacionadas con la ejecución remota y la gestión de archivos y directorios críticos. `MailServer`, basado en Linux, presenta características similares en cuanto a acceso remoto y almacenamiento de información.

El dispositivo de seguridad en la red es un *firewall*, que cumple una función de protección. Cuenta con permisos administrativos y la capacidad de capturar tráfico de red y analizar paquetes, lo que permite monitorear y restringir accesos no autorizados.

También se ha modelado una aplicación, `Apache`, que opera en Linux con permisos administrativos. Esta aplicación tiene acceso a archivos y directorios específicos y presenta vulnerabilidades potenciales debido a software no actualizado.

En cuanto a las relaciones, `Alice` se autentica en su equipo, el cual accede a `WebServer` mediante el protocolo LDAP, mientras que el equipo utilizado por `Bob` establece comunicación con `MailServer` utilizando el protocolo SMTP. Además, `WebServer` y `MailServer` están enlazados mediante una conexión en el puerto 25 con protocolo SMTP, lo que indica la comunicación entre estos servicios.

El nodo `Firewall` protege tanto a `WebServer` como a `MailServer`, reforzando la seguridad de la red. Finalmente, `Apache` está alojado en `WebServer`, lo que indica su dependencia de este servidor para operar dentro del entorno web.

Si ahora se ejecuta el comando `$ python3 argos.py attack`, el sistema generará una secuencia aleatoria de técnicas y evaluará su impacto sobre el escenario cargado. Al ejecutar de nuevo la *query* `MATCH (n) RETURN (n)`, se puede apreciar que la secuencia de ataque generada ha sido insertada y que los nodos y relaciones correspondientes han sido creados sobre el escenario (Figura 6).

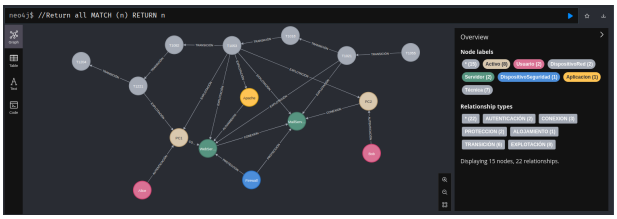


Figura 6. Ataque sobre el escenario de oficina corporativa

Se va a analizar una de las relaciones de explotación creadas, en concreto la que asocia a la técnica T1021 (uno de los nodos de color gris en la Figura 6) con el activo `WebServer` (representado en verde).

Lo primero es verificar que los nodos se han creado correctamente, conteniendo los atributos que los caracterizan. El nodo `MailServer` representa el servidor de correo dentro de la infraestructura de red, es un nodo de tipo `Activo` y subtipo `Servidor`, sus atributos se muestran en la Figura 7.

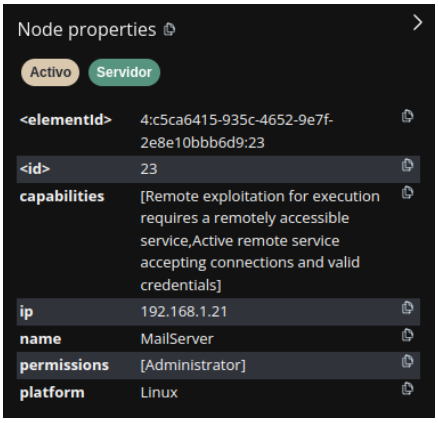


Figura 7. Propiedades del nodo `MailServer`

Por otro lado, el nodo `T1021`, de tipo `Técnica`, representa la técnica *Remote services* de la matriz empresarial de MITRE ATT&CK. Esta técnica describe cómo los adversarios pueden usar cuentas válidas para acceder remotamente a sistemas dentro de una red empresarial que cuenta con un sistema de inicio de sesión centralizado. Si logran credenciales legítimas, los atacantes pueden iniciar sesión en múltiples dispositivos mediante protocolos como SSH o RDP, o incluso en servicios en la nube vinculados al dominio. Sus atributos se pueden apreciar en la Figura 8.

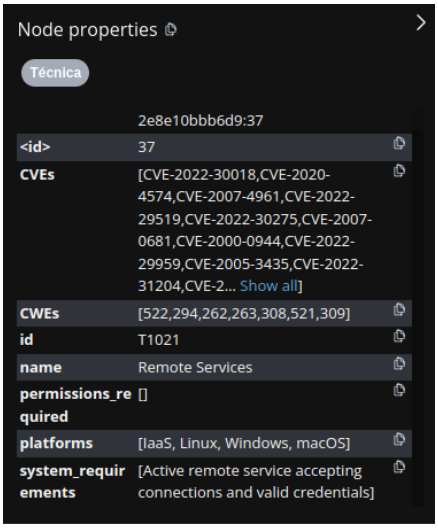


Figura 8. Propiedades del nodo `T1021`

Al no haber ninguna vulnerabilidad configurada en el atributo `cve` de `MailServer`, el origen de la relación de `EXPLOTACIÓN` han de ser los parámetros indicados en naranja en la Figura 3. Si se revisan, se puede apreciar que la plataforma que corre el servidor, Linux, es una de las plataformas afectadas por el sistema. Además, el sistema cuenta con la capacidad *Active remote service accepting connections and valid credentials*, uno de los requisitos necesarios por esta técnica para ser explotada. Finalmente, la técnica no requiere de ningún permiso en concreto para ser explotada con éxito. De hecho, esta técnica se clasifica en la matriz empresarial dentro de la táctica *Lateral Movement*, pues al obtener credenciales válidas de dominio, un atacante puede desplazarse dentro de la red accediendo a múltiples dispositivos y servicios sin necesidad de explotar vulnerabilidades adicionales.

Se cumplen, por tanto, las tres condiciones establecidas para que exista una relación de tipo `EXPLOTACIÓN` entre estos dos nodos.

A la hora de asociar la secuencia de ataque generada con los activos del escenario, el comando `attack` extrae simultáneamente una serie de métricas acerca del ataque que ha tenido lugar. Estas estadísticas se presentan al terminar la ejecución a través de la terminal, organizadas en paneles de forma clara y jerárquica mediante la librería Rich. Este *dashboard* permite a los usuarios interpretar de manera rápida y eficaz los resultados del ataque simulado. La información mostrada es la siguiente, y se puede observar en la Figura 9.

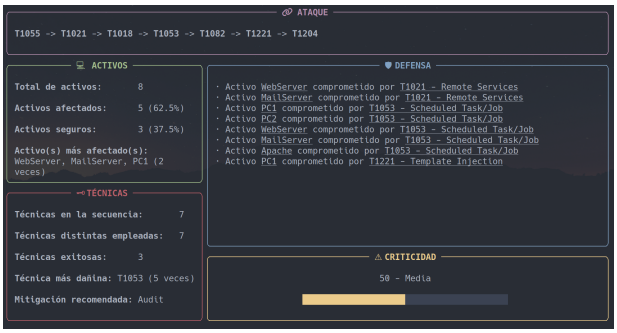


Figura 9. Resumen del ataque presentado sobre el escenario de oficina

- Panel `ATAQUE`: muestra la cadena de técnicas empleadas en el ataque. Las técnicas aparecen en orden y conectadas con flechas, lo que permite visualizar claramente cómo se ha desarrollado el ataque desde su inicio hasta su última fase.
- Panel `DEFENSA`: muestra una lista detallada en la que se indica, para cada activo comprometido, la técnica específica que lo ha vulnerado. Este panel proporciona información valiosa para identificar puntos débiles en la infraestructura y planificar medidas correctivas.
- Panel `ACTIVOS`: este panel proporciona una serie de estadísticas sobre los activos del escenario, permitiendo tener una visión clara del impacto del ataque sobre la infraestructura tecnológica. En concreto, se muestran las siguientes métricas:
  - `Total de activos`: número total de activos presentes en el escenario cargado.
  - `Activos afectados`: número y porcentaje de activos que han sido comprometidos por, al menos, una técnica.
  - `Activos seguros`: número y porcentaje de activos que no han sido vulnerados.
  - `Activo(s) más afectado(s)`: de entre los activos afectados, activo o activos comprometidos por el mayor número de técnicas.
- Panel `TÉCNICAS`: este panel resume datos sobre las técnicas empleadas, lo que permite enfocar las medidas defensivas en los métodos de ataque más dañinos.

  - `Técnicas en la secuencia`: número total de técnicas en la secuencia.
  - `Técnicas distintas empleadas`: de entre las técnicas totales de la secuencia, número de técnicas que aparecen una única vez en ella.
  - `Técnicas exitosas`: número de técnicas que han logrado comprometer, al menos, un activo.
  - `Técnica más dañina`: identificador de la técnica que ha comprometido el mayor número de activos distintos.
  - `Mitigación recomendada`: estrategia de defensa recomendada para paliar el impacto de la técnica más dañina.
- Panel `CRITICIDAD`: utilizando los datos anteriores, se le asigna una criticidad (`C`) de entre 0 y 100 al ataque, calculada como:

$$C = \left(\left\lfloor \frac{\text{Total activos afectados}}{\text{Total activos}} \times 50 \right\rfloor + \left\lfloor \frac{\text{Total técnicas exitosas}}{\text{Total técnicas empleadas}} \times 50 \right\rfloor \right)$$

  Esta criticidad se visualiza mediante una barra de progreso coloreada de la siguiente forma:
  - Verde (baja) si $C \leq 33$
  - Amarillo (media) si $33 < C \leq 66$
  - Rojo (alta) si $66 < C \leq 100$

VII. Conclusiones y Líneas Futuras

El análisis de las rutas de ataque es uno de los puntos clave en la investigación de ciberseguridad actualmente. Para ello, la visualización del camino que este ataque puede recorrer dentro de una red resulta una herramienta eficaz para el estudio de los ciberataques con el objetivo de prevenirlos o mitigar su efecto dentro de entornos organizacionales. Al proporcionar una representación gráfica de los posibles recorridos de un atacante dentro de una red o sistema, se contribuye a la identificación de vulnerabilidades y la toma de decisiones estratégicas en seguridad, objetivos clave de los entornos de conciencia cibersituacional.

Los resultados obtenidos con ARGOS se complementan con el desarrollo de una estrategia de seguridad que permita abordar la respuesta frente a las amenazas a raíz de la caracterización de las técnicas utilizadas. El análisis de las rutas de ataque desde el punto de vista planteado en ARGOS permite identificar activos vulnerables, el impacto de los ataques y las contramedidas más eficaces; mientras que el estudio de los ataques para su caracterización, en este contexto, permite definir modelos más adaptados al comportamiento de los atacantes y que generen simulaciones más realistas.

Con el objetivo de mejorar la funcionalidad y el alcance de ARGOS, se presentan a continuación diversas líneas de desarrollo que podrían implementarse en versiones futuras de la herramienta. La implementación de estas mejoras permitiría que ARGOS se convierta en una herramienta más robusta y adaptable a las necesidades de análisis de seguridad en entornos de red.

Actualmente, ARGOS permite la carga de escenarios de red definidos mediante ficheros Cypher, pero la complejidad de estos escenarios podría ampliarse. Se podría considerar la integración de configuraciones más detalladas, incluyendo topologías de red dinámicas, segmentación de redes, o la representación de defensas activas como sistemas de detección de intrusiones (IDS) y *firewalls*.

La simulación de ataques mediante Cadenas de Markov podría beneficiarse de una mayor personalización en la generación de secuencias de TTPs. Por ejemplo, podría incorporarse un sistema de pesos dinámicos basado en inteligencia de amenazas en tiempo real, permitiendo que las probabilidades de transición entre técnicas cambien en función de tendencias recientes en ataques reales. Otra posible mejora sería la incorporación de modelos probabilísticos más avanzados, como redes bayesianas, para mejorar la realismo de las transiciones entre estados.

En el estado actual, la herramienta almacena métricas sobre los ataques ejecutados, pero se podría ampliar la capacidad de análisis mediante la generación de informes automáticos con visualizaciones interactivas. El uso de librerías de análisis de datos como Pandas o herramientas de visualización como Dash podría permitir una exploración más intuitiva del impacto de los ataques en los activos del sistema.

Por último, el almacenamiento del historial de ataques podría enriquecerse con un sistema de versionado que permitiera comparar la evolución de escenarios a lo largo del tiempo. Además, se podría incluir un módulo de aprendizaje automático para identificar patrones en ataques previos y predecir posibles rutas de ataque futuras.

Referencias

[1] Check Point: Cyberattacks targetting governments https://www.checkpoint.com/es/cyber-hub/cyber-security/what-is-cybersecurity-for-governments/cyberattacks-targeting-governments/. Último acceso: 17 de febrero de 2025.

[2] Telefónica Tech: Tendencias en Ciberseguridad para 2025 https://telefonicatech.com/blog/tendencias-en-ciberseguridad-para-2025. Último acceso: 6 de febrero de 2025.

[3] Plataforma Tecnológica Española de Tecnologías Disruptivas https://ptedisruptive.es/wp-content/uploads/2024/12/INFORME-DE-SITUACION_CIBERSEGURIDAD_2024.pdf. Último acceso: 6 de febrero de 2025.

[4] Q. Wang, W. Wang, Y. Wang, J. Ren, B. Zhang: "Multi-Stage Network Attack Detection Algorithm Based on Gaussian Mixture Hidden Markov Model and Transfer Learning" en *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 3470-3484, 2025.

[5] Pilar Holgado, Víctor A. Villagrá, Luis Vázquez: "Real-Time Multistep Attack Prediction Based on Hidden Markov Model", en *IEEE Transactions on Dependable and Secure Computing*, vol. 17, n. 1, pp. 134-147, 2017.

[6] M. A. R. A. Amin, S. Shetty, L. Njilla, D. K. Tosh, C. Kamhoua: "Hidden Markov Model and Cyber Deception for the Prevention of Adversarial Lateral Movement", en *IEEE Access*, vol. 9, pp. 49662-49682, 2021.

[7] S. Dass, P. Datta, A. S. Namin: "Attack Prediction using Hidden Markov Model", en *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1695-1702, 2021.

[8] Fuxiong Sun, Juntao Pi, Jin Lv, Tian Cao: "Network Security Risk Assessment System Based on Attack Graph and Markov Chain", en *Journal of Physics: Conference Series*, 910, 2017.

[9] Ricardo M. Czekster, Charles Morisset: "BDMPathfinder: a tool for exploring attack paths in models defined by Boolean logic Driven Markov Processes", en *2021 17th European Dependable Computing Conference (EDCC)*, pp. 83-86, 2021.

[10] The Mitre Corporation: MITRE ATT&CK https://attack.mitre.org/. Último acceso: 10 de enero de 2025.

[11] Universidad Politécnica de Madrid: Cadenas de Markov https://dcain.etsin.upm.es/~carlos/bookAA/04.02_CadenasMarkovResultados.html. Último acceso: 10 de marzo de 2025.

[12] Becca Gomby: What Is An Attack Path & How Does It Help Identify Risks? https://www.panoptica.app/blog/what-is-an-attack-path-how-does-it-help-identify-risks. Último acceso: 30 de enero de 2025.

[13] Picus Labs: What Is Attack Path Visualization? https://www.picussecurity.com/resource/glossary/what-is-attack-path-visualization. Último acceso: 30 de enero de 2025.

(Work in Progress) Automated Synthesis of Instruction-Centric Leakage Contracts

Elvira Moreno Sánchez †
IMDEA Software Institute
Universidad Politénica de Madrid (UPM)
elvira.moreno@imdea.org.es

Tiziano Marinaro †
CISPA Helmholtz Center for Information Security
Saarland University
tiziano.marinaro@cispa.de

Ryan Williams
Northeastern University
williams.ry@northeastern.edu

Marco Patrignani
University of Trento
marco.patrignani@unitn.it

Hamed Nemati
KTH Stockholm
hnnemati@kth.se

Roberto Guanciale
KTH Stockholm
robertog@kth.se

Marco Guarnieri
IMDEA Software Institute
marco.guarnieri@imdea.org

***Abstract*—Side-channel attacks pose a significant security threat for modern computing platforms, because they exploit subtle discrepancies in CPU behaviors to leak sensitive information. To model the information leaked by a CPU via microarchitectural side-channels, recent work proposed *leakage contracts*: a security abstraction that provides the foundations for secure CPU programming. Unfortunately, due to the complexity of current microarchitectural optimizations, devising leakage contracts requires extensive manual effort and thus modern CPUs lack dedicated leakage contracts.**

In this paper, we present our on-going work towards an automated methodology to extract sound and precise instruction-centric leakage contracts for two major CPU architectures with the minimum degree of manual intervention. We implemented this methodology in the MALCOS tool, which automates the synthesis of leakage contracts for black-box CPUs. We evaluate MALCOS on `x86` and `ARM` CPUs, and show that the contracts it synthesizes are sound and precise. Our results demonstrate that extracting leakage contracts for modern CPUs is feasible even when their leakage templates are not available.

***Index Terms*—Microarchitectural attacks, Leakage contracts, Program synthesis**

Type of contribution: *Research in progress (limit 8 pages)*

I. Introduction

Microarchitectural attacks [1], [2] exploit subtle differences in a program's execution time—due to CPU optimizations—to leak information from seemingly secure programs. To defend against such attacks, programmers need to reason about the interactions between software and a CPU's microarchitecture. However, the Instruction Set Architecture (ISA)—the traditional abstraction layer between software and hardware—lacks microarchitectural details.

Leakage contracts [3] augment the Instruction Set Architecture (ISA) with a specification of all observable side-channel leaks associated with a given CPU. This enables secure system development since programmers are made aware of the exploitable side-channels that are traditionally obscured at the ISA level. Unfortunately, constructing leakage contracts for modern CPUs is a complex task: it requires extensive reverse engineering, expert knowledge, and significant time investment [4], making it impractical to apply across the diverse landscape of commercially available CPUs. The lack of a scalable, automated method for generating accurate leakage contracts limits their applicability and effectiveness in real-world security.

In this work, we report on our on-going work to extract leakage contracts from black-box CPUs. Our methodology focuses on *instruction-centric* leakage contracts, i.e., contracts where leaks are characterized as a function of instruction operands, which can capture subtle data-dependent instruction-level optimizations [5]. We make the following contributions:

Domain-Specific Language (DSL): We formalize leakage contracts using a Domain-Specific Language (DSL) (Section III). Following the formalism from [3], our DSL models leakage contracts as a set of *leakage clauses* describing what observations are leaked during program execution. That is, leakage contracts in out DSL map each program execution to a *leakage trace* capturing the leaked information. Importantly, contracts specified in our DSL are *executable*. That is, they can be applied to arbitrary `x86`, `ARM` programs (and associated inputs) to derive the corresponding *leakage traces* recording all leaks modeled by the contract.

Contract synthesis methodology: We develop a novel methodology, which we overview in Section II and detail in Section IV, for automatically synthesizing instruction-centric leakage contracts based on hardware observations extracted from a black-box CPU. For this, we employ a counterexample-driven synthesis method that iteratively refines candidate leakage contracts based on observed hardware behavior. Our syntesis approach incrementally learns the leakage contract associated with a CPU by (1) given a candidate leakage contract (intially empty) capturing the leaks observed so far, we use existing relational leakage testing tools [6], [7] to validate the contract against a black-box CPU, (2) newly discovered leaks are synthesized into a new ISA-level contract clause. These clauses are then added to the original candidate contract, thereby ensuring that the candidate captures the discovered leak.

Once no more clauses can be added to the contract, the contract is minimized to remove redundant clauses. Our methodology ensures that the synthesized contract is *sound*, i.e., it captures *all* leaks exposed during testing. Furthermore, learned contracts have *high precision*, i.e., they do not unnecessarily over-approximate actual leaks.

†These authors contributed equally to this work

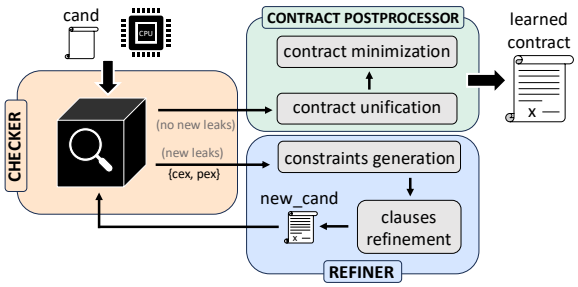


Fig. 1: MALCOS takes as input a candidate contract and a CPU, and it uses a combination of fuzzing and synthesis to refine clauses that capture discovered leaks. When the CHECKER cannot capture more leaks, MALCOS minimizes the contract and outputs the final result.

MALCOS synthesis tool: We implement our methodology in a tool called MALCOS (MicroArchitectural Leakage COntract Synthesizer) (Section V). MALCOS incrementally builds an instruction-centric leakage contract capturing leaks in a given black-box CPU, and it is applicable across different architectures and hardware configurations. In particular, it supports both `x86` and `ARM` CPUs.

Unlike existing work [8], [9], [10], MALCOS eliminates the need for predefined contract templates or manual test-case generation (as required by [9]). Moreover, it does not require having access to the processor Register Transfer Level design (like [10]), and it is not limited to coarse-grained contracts classifying instructions into safe or unsafe (like [8]).

Our preliminary results (Section VI) indicate that MALCOS can successfully synthesize program-level leakage contracts, showing its ability to identify a sound over-approximation of standard leakage models.

II. OVERVIEW

In this section, we overview MALCOS's synthesis approach with an example. At its core, our methodology relies on the components depicted in Figure 1 in order to implement the leakage contract synthesis.

a) CHECKER*:* The CHECKER takes a candidate contract `cand` and a CPU (that is treated like a black-box) and it tries to discover leaks in the target CPU that are not already captured by `cand`. Whenever the CHECKER finds a new leak, it returns a *counterexample* (`cex`), which is a sequence of instructions together with a pair of initial states that produce the *same leakage traces* according to the candidate contract, but result in different microarchitectural observations, i.e., they are distinguishable by a microarchitectural attacker. Additionally, the CHECKER also returns *positive examples* (`pex`), i.e., test cases that are indistinguishable for both the contract and the microarchitectural attacker. These are used to correct contract over-approximation, as we discuss in Example 2 below.

Our methodology is centered on contracts because of the notion of *contract satisfaction* [3], which precisely characterize whenever a contracts capture all leaks in the underlying CPU. We rely on this notion in order to determine whether a pair of test cases form a counterexample or a positive example.

Our methodology is parametric in the underlying CHECKER and in fact MALCOS uses REVIZOR [6] and SCAM-V [7] (both using cache-based side-channels as a source of microarchitectural observations) to detect leaks in `x86` and `ARM` processors respectively.

b) REFINER*:* The REFINER takes a counterexample `cex` describing a newly-discovered leak and it generates a new contract clause that captures the new leak. This clause is then added to the original contract in order to generate a new candidate contract `new_cand`. The REFINER discovers such a contract clause as a syntax-driven synthesis task implemented on top of the Rosette solver [11].

c) CONTRACT POSTPROCESSOR*:* The CONTRACT POSTPROCESSOR takes a candidate contract and it tries to minimize the contract by unifying different clauses and by reducing the number of unnecessary clauses. The goal of the CONTRACT POSTPROCESSOR is to reduce the complexity of the final contract (`learned contract`).

To showcase what a synthesized contract is, and the synthesis process itself, Example 1 presents an application of our methodology. We use this example to introduce a key contract synthesis issue: over-approximating contracts, and our solution to it in Example 2. Examples 1 and 2 consider a CPU that implements a simple *register file compression* (RFC) optimization [5]. This optimization reduces the physical size of the register file by mapping all logical registers that store the value 0 to the same physical register. These compression schemes, however, often reduce the pressure on the register file thereby potentially resulting in timing leaks [5], [12].

Example 1 (RFC Synthesis)**.** Starting from an empty candidate contract ($\texttt{cand} = \emptyset$) the CHECKER must identify any leaks. For this, the CHECKER generates test cases, each one consisting of a program and a pair of initial states, executes them on our target CPU, and performs microarchitectural observations[1] to detect potential leaks. Given that $\texttt{cand} = \emptyset$, while the target CPU leaks through RFC, the CHECKER discovers the following counterexample describing the RFC leak:

$$p := \texttt{MOV rax,rbx} \quad \sigma_1 := (\texttt{rbx} \mapsto 0) \quad \sigma_2 := (\texttt{rbx} \mapsto 1)$$

Here, the program p consists of an instruction assigning to register `rax` the value of register `rbx`, where the value of `rbx` is 0 in initial state σ_1 and a value different from 0 in initial state σ_2. Therefore, RFC happens when executing p from σ_1, but does not happen when executing p from σ_2, which results in different microarchitectural observations for the attacker. That is, the tuple $\texttt{cex} := \langle p, \sigma_1, \sigma_2 \rangle$ is a counterexample for the candidate contract $\texttt{cand} = \emptyset$. The CHECKER ends by simulating the architectural execution of the counterexample and collects information about all architectural states (e.g., register values, operand information for the executed instructions, the program counter) explored during execution.

Next, the REFINER analyzes the counterexample $\texttt{cex} := \langle p, \sigma_1, \sigma_2 \rangle$ to generate a contract clause capturing the leak.

[1]For simplicity, both here and in Example 2, we consider a microarchitectural observer that sees whenever RFC happens.

The REFINER first generates the clauses for the synthesis problem from the given architectural states σ_1 and σ_2. Then the REFINER refines these clauses by using the Rosette solver to identify a new contract clause that distinguishes the counterexample `cex`. This yields the contract clause cl_1 below, which exposes the final value of the first operand (index 0) whenever the first operand is a register (type is `reg`) and it is the target of a register write (access is `write`). Finally, the REFINER adds the newly-discovered clause cl_1 to the candidate contract `cand`.

$cl_1 :=$ *post-operand-value 0* `IF` [(*operand-type 0* = `reg`) `and` (*operand-access 0* = `write`)]

This process is iterated over, performing a new round of checking and refinement to discover further leaks ignored by the candidate contract `cand` until no further leaks are found. When the candidate contract is sufficient to capture the RFC leaks, MALCOS uses the CONTRACT POSTPROCESSOR to return a simpler and more precise version of the contract. Given that cl_1 is sufficient to capture the RFC leaks (since it exposes all values written to registers during execution), and it only contains one single clause, MALCOS terminates and outputs the learned contract cl_1.

Example 2 (RFC Synthesis Made Precise)**.** During its iterative process, our methodology may synthetize an over-approximating contract, i.e., a contract that exposes more information than what is needed in order to capture the microarchitectural leaks. Such over-approximations are dangerous, since they can result in the synthesis of incorrect contract clauses.

For instance, consider the learned contract cl_1, which exposes the value of written registers during program execution. While cl_1 distinguishes any two executions leaking through RFC, it would also distinguish executions where no RFC happens, e.g., where no register takes the value 0. That is, cl_1 over-approximates the CPU's leakage, as it describes *more* leakage than the actual CPU.

To demonstrate this over-approximation, consider the same program p described above, and a *different* pair of states σ_3 and σ_4 where the values of register `rbx` are both different from 0.

$p :=$ `MOV rax,rbx` $\quad \sigma_3 := (\texttt{rbx} \mapsto 1) \quad \sigma_4 := (\texttt{rbx} \mapsto 2)$

Executing p in these states results in indistinguishable executions for our RFC attacker, i.e., there are no leaks to be found from these states. Yet, the previously synthesized contract cl_1 distinguishes these executions and thus over-approximates the desired behavior.

To address these over-approximations, our synthesis methodology relies on tuples such as the one above, which we dub *positive examples*, i.e., test cases that are indistinguishable for both the contract and the microarchitectural attacker. These positive examples are used to tell the synthesis methodology which executions should *not* be different, thus positive examples play a dual role to counterexamples. The goal of our synthesis methodology is now to synthesize a contract clause that distinguishes *as many counterexamples* while distinguishing *as few positive examples* as possible. Concretely, using positive examples, our methodology can learn the contract clause cl_2 below, which exposes when the first operand is a register and it is written with a value `0`.

$cl_2 :=$ *BOOL True* `IF` [(*operand-type 0* = `reg`) `and` (*operand-access 0* = `write`) `and` (*post-operand-value 0* = `0`)]

Crucially, clause cl_2 is *more precise* than cl_1, as it still captures RFC leaks because it only exposes whenever `0` is written to a register rather than exposing all register values throughout the execution.

Having clarified our methodology, this paper now provides formal details of its key element: contracts.

III. LEAKAGE CONTRACTS

This section formalizes the syntax (Section III-A) and the semantics (Section III-B) of contracts, as part of a language we dub ICL. Finally, we instantiate the notion of *contract satisfaction* [3] (Section III-C), which is used to detect counterexamples and positive examples.

A. ICL Syntax

(Bitstrings)	bs	$:=$	$\{0,1\}^n$
(Expressions)	ex	$:=$	$bs \mid \ominus ex \mid ex_1 \oplus ex_2 \mid ex[ex_1 : ex_2]$
		$\mid$	$\texttt{REG}(bs) \mid \texttt{POST_REG}(bs)$
		$\mid$	$\texttt{OP_TYPE}(bs) \mid \texttt{OP_ACC}(bs)$
		$\mid$	$\texttt{OP_VAL}(bs) \mid \texttt{POST_OP_VAL}(bs)$
		$\mid$	$\texttt{OPCODE}(bs)$
(Predicates)	pr	$:=$	$\{0,1\} \mid ex_1 = ex_2 \mid ex_1 < ex_2$
		$\mid$	$\texttt{NOT}\ pr \mid pr_1\ \texttt{AND}\ pr_2 \mid pr_1\ \texttt{OR}\ pr_2$
(Clause)	cl	$:=$	$ex\ \texttt{IF}\ pr$
(Contract)	C	$:=$	$\emptyset \mid C, cl$

Fig. 2: ICL syntax

The syntax of ICL is given in Figure 2. In ICL, a *contract* consists of a set of leakage clauses (using the terminology of [3], [12]), where each clause describes a piece of information that might be leaked through side channels by the underlying CPU. Each clause cl consists of a conditional expression ex `IF` pr where the expression ex indicates the leaked information and the predicate pr expresses under which conditions the leak happens.

The basic types of ICL are bit strings bs, with bitvectors of length one representing booleans. Expressions ex are constructed by combining bitstrings bs with unary $\ominus ex$ and binary $ex_1 \oplus ex_2$ operators, as well as the slice operator $ex[ex_1 : ex_2]$.

Additionally, expressions can refer to the registers ($\texttt{REG}(bs)$), the opcode of the instruction ($\texttt{OPCODE}(bs)$), or information about the operands, such as the type ($\texttt{OP_TYPE}(bs)$), its access ($\texttt{OP_ACC}(bs)$), or its value ($\texttt{OP_VAL}(bs)$).

Moreover, expressions may cover operands and registers in the next architectural state, i.e., after the current instruction has been executed ($\texttt{POST_REG}(bs)$ and $\texttt{POST_OP_VAL}(bs)$).

Finally, predicates can be boolean values, equalities and inequalities between expressions, or combinations of predicates with the standard logical operators.

Program-level Contracts: In general, leakage contracts are defined for all possible programs of a CPU architecture.

However, they can also be specified for a single program by constraining them to only contain clauses ex IF pr where pr constrains the program counter to be a constant (pointing to one of the program instructions).

B. ICL Semantics

The key element of interest in the semantics of leakage contracts is *leakage* traces, i.e., the sequence of observations obtained by evaluating the clauses in the contract on *every state* in the execution. In order to collect these traces (Section III-B3), ICL requires the definition of a system semantics, which captures how the underlying system works (Section III-B1) and of a contract semantics, which dictates how the single observations that constitute a leakage trace are generated (Section III-B2).

1) System Semantics: The system semantics $\rightarrow$: $\sigma \times \sigma$ maps each state to the next one by executing one instruction in program p (written $p \vdash \sigma \rightarrow \sigma'$). The system semantics is left abstract except for these requirements. First, the set $\Sigma_0 \subseteq \Sigma$ denotes the set of initial states, where Σ is the set of all states. Second, states σ contain a memory m and a register assignment r that, respectively, represent the state of memory and registers. Memories m map addresses (represented as integers) to values. Register assignments r map register identifiers (represented as bitstrings) to their values. We also assume that one of the register identifiers, denoted PC, represents the program counter register.

2) Contract Semantics: The semantics of an ICL contract C is obtained by extending the system semantics with observations generated according to C. Formally, each ICL contract C induces a labeled semantics $\rightarrow_C$: $\sigma \times \delta \times \sigma$ specified by this inference rule:

$$\frac{p \vdash \sigma \rightarrow \sigma' \quad \delta = \{[\![ex]\!](\sigma,\sigma') \mid ex \text{ IF } pr \in C \wedge [\![pr]\!](\sigma,\sigma') = \top\}}{p \vdash \sigma \xrightarrow{\delta}_C \sigma'}$$

where $[\![ex]\!](\sigma,\sigma')$ (respectively, $[\![pr]\!](\sigma,\sigma')$) denotes the observation obtained by evaluating expression ex (respectively, predicate pr) in states σ and σ'. Note that the evaluation function $[\![\cdot]\!]$ needs also the next-state σ' to evaluate expressions of the form $\texttt{POST_REG}(bs)$ and $\texttt{POST_OP_VAL}(bs)$. In a nutshell, the rule above executes one step of the system semantics ($p \vdash \sigma \rightarrow \sigma'$) and labels it with the set δ of leakage observations obtained by evaluating in the current state the expressions of all leakage clauses whose predicates are satisfied ($\delta = \{[\![ex]\!](\sigma,\sigma') \mid ex \text{ IF } pr \in C \wedge [\![pr]\!](\sigma,\sigma') = \top\}$).

3) Leakage Traces: A *leakage trace* τ is a sequence of observations obtained by applying the contract semantics $\rightarrow_C$ to an execution $p \vdash \sigma_0 \rightarrow \sigma_1 \rightarrow \ldots$, that is, $\tau = \delta_0\delta_1\ldots$ where $p \vdash \sigma_0 \xrightarrow{\delta_0}_C \sigma_1 \xrightarrow{\delta_1}_C \ldots$ In the following, we write $\texttt{CTR}(C,p,\sigma)$ to denote the leakage trace (under contract C) associated with the maximal execution of the program p starting from the initial state σ. Furthermore, we say that two initial states σ, σ' are *C-equivalent* for a program p whenever they result in the same leakage traces, $\texttt{CTR}(C,p,\sigma) = \texttt{CTR}(C,p,\sigma')$

C. Contract Satisfaction

Contract satisfaction precisely characterizes under which conditions a contract captures all leaks in an underlying CPU.

Following [3], [6], we first introduce the notion of hardware traces, which capture the observational power of the (microarchitectural) attacker operating at the level of the underlying CPU. We represent the hardware trace as the output of the function $\texttt{HTR}(p, \sigma, Ctx)$, that returns the observations made by an attacker on all available microarchitectural side channels. The function $\texttt{HTR}(\cdot)$ takes as inputs the victim program p, the initial state σ processed by the victim program (i.e., the architectural state including registers and main memory), and the microarchitectural context Ctx in which it executes (i.e., the initial state of microarchitectural components—like caches and predictors—in the underlying CPU).

Informally, a CPU *satisfies a leakage contract C for a program p* [3] if any two states that are C-equivalent also result in identical hardware traces in any context, i.e., the executions are indistinguishable by a microarchitectural attacker. Formally:

$$p \vdash C \stackrel{\text{def}}{=} \forall \sigma, \sigma', Ctx. \\ \text{if } \texttt{CTR}(C,p,\sigma) = \texttt{CTR}(C,p,\sigma') \text{ then } \texttt{HTR}(p,\sigma,Ctx) = \texttt{HTR}(p,\sigma',Ctx)$$

We say that a CPU *satisfies a contract C*, if the CPU satisfies it for every program. Dually, if a CPU does not satisfy a contract C, there is a *leakage counterexample* $\texttt{cex} := \langle p, \sigma, \sigma' \rangle$ consisting of a program p and two initial states σ, σ' that produce different hardware traces (for some context Ctx) but it generates identical leakage traces for C.

IV. Synthesizing Leakage Contracts

In this section, we present our methodology for automatically synthesizing leakage contracts, outlined in Algorithm 1.

Our algorithm automatically constructs a leakage contract from a CPU under test (denoted $target$ in this section) in a black-box manner. That is, the synthesis algorithm does not have access to the CPU design, and it only interacts with the CPU by executing *test cases*, each one consisting of a program p and a sequence of initial states, in order to derive the corresponding hardware traces.

Algorithm 1 describes the contract synthesis algorithm, which takes two inputs: the CPU under test $target$ and the number of test programs n used in the synthesis. The algorithm works in two phases, which we describe next.

Phase 1 spans lines 1–14, and it synthesizes a set of leakage contracts that capture the leaks explored by n test programs. In this phase, the algorithm generates *program-level* ICL contracts. This allows us to capture fine-grained leaks that may depend on a specific instruction context.

Phase 1 starts by computing a set P of n (randomly generated) test programs (line 2). For each program $p \in P$, the algorithm first computes the starting initial candidate $cand$ (line 6). This candidate can either be $\emptyset$ or derived from the contracts generated from prior test programs.

Next, the algorithm checks whether there are leaks in the underlying CPU under test $target$ that are not captured yet by the candidate contract $cand$ (line 8). For this, the algorithm relies on the CHECKER function, which returns a set of counterexamples $cexs$ and a set of positive examples $pexs$ for program p. The former are leaks not yet captured by the contract $cand$, whereas the latter represent executions that are indistinguishable for the microarchitectural attacker (which we will use to improve the precision of the synthesized

Algorithm 1: MALCOS contract synthesis approach

```
1  Procedure ContractSynthesis(target, n):
       // Phase 1 - program-level synthesis
2      P ← generateSeedPrograms(n);
3      cands ← [];
4      CEXs ← [];
5      for p ∈ P do
6          cand ← getInitialCtr(cands);
7          while true do
8              cexs, pexs ← Checker(p, cand, target);
9              CEXs[p] ← CEXs[p] ∪ cexs;
10             if cexs ≠ ∅ then
11                 cand ← Refiner(p, cand, cexs, pexs);
12             else
13                 break
14         cands ← append(cands, cand);
       // Phase 2 - unification and minimization
15     candCtr ← unify(cands);
16     ctr ← minimize(candCtr, CEXs);
17     return ctr;
```

contract). If CHECKER finds at least one counterexample, the algorithm refines the current candidate contract $cand$ to account for the newly discovered leaks using the `Refiner` function (line 11), and it later restarts the loop. Otherwise, if CHECKER cannot find any further counterexamples, the algorithm terminates the testing of the current program, appends the candidate contract $cand$ to the sequence $cands$ (line 14), and moves to the next program in P.

Phase 2 spans lines 15–17, and it derives the final leakage contract for the target CPU by generalizing the program-level contracts generated during Phase 1. First, this phase unifies all program-level contracts obtained so far into a single ICL candidate contract using the `unify` function (line 15). Importantly, this step produces a contract that is *architecture-level*, i.e., predicates in the leakage clauses of this contract are not tied to specific program counters but rather to instruction types. Next, the algorithm minimizes the architecture-level leakage contract using the `minimize` function (line 16) and returns it to the user. The minimization step allows us to remove unnecessary leakage clauses from the contract, while preserving soundness w.r.t. the explored counterexamples. That is, the final leakage contract still distinguishes all counterexamples in $CEXs$, and it is more precise (i.e., it distinguishes fewer pairs of states with the same hardware trace) than the non-minimized contract.

V. IMPLEMENTATION

This section presents the MALCOS contract synthesis tool, which consists of two separate toolchains (called MALCOS-X86 and MALCOS-ARM) to cater for the `x86` and `ARM` architectures. Both toolchains follow the workflow outlined in Algorithm 1, but they differ in the implementation of: the contract satisfaction tool (Section V-A), and the way contracts are refined (Section V-B).

A. Code Examples and Contract Satisfaction

The CHECKER part of MALCOS generates test programs (function `generateSeedPrograms` in Algorithm 1) and extract counter- and positive examples for a given contract candidate (function `Checker` in Algorithm 1). A *counterexample* consists of two $cand$-equivalent states σ, σ' with different hardware traces, i.e., a microarchitectural attacker can distinguish the executions of p starting from σ and from σ' through a microarchitectural side-channel. In contrast, a *positive example* consists of two $cand$-equivalent states σ, σ' that additionally have the same hardware traces, i.e., two states that are indistinguishable for a microarchitectural attacker.

MALCOS implements the CHECKER with state-of-the-art tools for relational contract testing: REVIZOR [6] for `x86` and SCAM-V [7] for `ARM`. In a nutshell, both tools take as input a contract and try to discover contract violations (i.e., counterexamples) in the CPU under test by generating random programs and inputs, executing them both on the actual CPU and at contract-level, and comparing contract and hardware traces. Off-the-shelf, both tools only output counterexamples, so we modified them to extract positive examples as well. It is worth noting that each call to the `Checker` produces a single counterexample in MALCOS-X86 but yields multiple counterexamples in MALCOS-ARM. We will provide further details on the two toolchains next.

a) MALCOS-X86*:* For `x86`, we rely on REVIZOR, which simulates test programs using the Unicorn CPU emulator extended with support for contracts. In MALCOS-X86, we extend REVIZOR to support ICL contracts by compiling them into event-handlers for the Unicorn emulator that computes leakage observations during execution [12].

Given that REVIZOR uses randomly-generated programs and inputs during the testing process, we implemented two optimizations to tame randomness and improve the precision of the synthesized contracts. The first optimization is *program minimization*, which minimizes test programs that result in violations by iteratively removing instructions from p so long as it can still find a counterexample in the smaller program p'.The second optimization is *data minimization*, which identifies those parts of the two architectural states σ and σ' causing a violation, and iteratively modifies σ' by copying part of the state from σ (at a byte-granularity) until the violation persists.

b) MALCOS-ARM*:* For `ARM`, we use SCAM-V instead of REVIZOR. Unlike REVIZOR, SCAM-V combines techniques from program verification and fuzzing to perform relational testing to validate the candidate contracts. For a generated test program p, SCAM-V uses *symbolic execution* to synthesize a relation that identifies which states are observationally equivalent according to the contract being validated. Next, SCAM-V generates an instance of this relation in terms of two input states. Finally, similar to REVIZOR, SCAM-V can run the generated program with two inputs on hardware to collect hardware traces.

B. Contract Refinement

To refine the candidate contract (function `Refiner` in Algorithm 1), we implement the clause synthesis process outlined in Section IV on top of Rosette [11], a solver-aided language that extends Racket. For this, we first formalize

ICL's syntax and semantics in Rosette. Next, we collect all sequences of states explored as part of the counterexamples $cexs$ and positive examples $pexs$. Rosette will then use the Z3 SMT solver [13] to synthesize an ICL leakage clause satisfying the constraints.

The main difference between the MALCOS-X86 and MALCOS-ARM toolchains is that MALCOS-X86 synthesizes one leakage clause per instruction in the program under test p during refinement, whereas MALCOS-ARM synthesizes a single leakage clause for the entire program. The reason for this difference lies in the program minimization optimization implemented in MALCOS-X86 (Section V-A). This optimization allows us to pinpoint *all* instructions involved in the leak, for each of which we synthesize a clause to capture the source of the leak. In contrast, MALCOS-ARM does not minimize the test program, which may include instructions not contributing to the leak. By synthesizing a single clause, we let the synthesis solver identify the leak sources (without introducing unnecessary approximations due to clauses associated with unrelated instructions).

VI. Evaluation

This section evaluates the ability of MALCOS to automatically learn leakage contracts for `x86` and `ARM` architectures. We aim to answer the following research questions:

RQ1: How good is MALCOS at synthesizing program-level contracts?

RQ2: How good is MALCOS at synthesizing architecture-level contracts?

Next, we first introduce the common metrics used to evaluate the quality of learned contracts (Section VI-A). Then, we describe the target leakage contracts used as ground truth in our experiments (Section VI-B). Finally, we report on **RQ1** and **RQ2** in Sections VI-C and VI-D.

A. Metrics

In the absence of standard measures, we define two metrics to assess the quality of the learned contracts: precision and soundness. These metrics are defined w.r.t. a *validation set* V, which consists of test cases (p, σ, σ') where p is a program and σ and σ' are initial states.

Precision measures how precisely the learned contract reflects leaks in the target system. Precision is defined as: $\frac{TP}{TP+FP}$, where TP is the number of true positives, i.e., test cases in V that are distinguishable using both contract and hardware traces, and FP is the number of false positives, i.e., test cases in V that are distinguishable only by the contract traces but not by the hardware traces. While traditionally used for binary classification, recent work [9] uses it for contract synthesis to discard fewer contract-level programs that could safely run on the processor.

Soundness measures the correctness of the learned contract. It is defined as $\frac{TP}{TP+FN}$, where now, FN is the number of false negatives, i.e., test cases in V that are distinguishable according to the hardware traces but not for the contract traces.

B. Contract Models

To answer **RQ1** and **RQ2**, we use MALCOS to synthesize leakage contracts against a set of fixed contract models used as ground truth, i.e., in MALCOS these models are used as *target* to obtain hardware traces. These models, described next, encapsulate representative microarchitectural behaviors and distinguish relevant leakage scenarios.

Contract	MALCOS-X86		MALCOS-ARM	
	avg. P/S	avg. Time	avg. P/S	avg. Time
ct	1.0/1.0	1min 48s	1.0/1.0	6m 32s
TagIdx	0.95/1.0	8min 19s	0.99/1.0	14m 29s
RFC	0.91/1.0	3min 16s	0.92/1.0	16m 52s
mul	0.94/1.0	1min 32s	0.92/1.0	24m 58s
SilStore	0.96/1.0	1min 19s	0.94/1.0	10m 33s

TABLE I: Progr.-level results: Precision (P) & Soundness (S)

- **Constant-time contract (ct):** models the constant-time observer [14], which is commonly used when reasoning about side channels in cryptographic algorithms. **ct** exposes the value of the program counter and the addresses of load and store operations throughout the execution.
- **Tag Index contract (TagIdx):** differs from **ct** in that it only exposes the tag and set index associated with each memory access (instead of the whole memory address like in **ct**). **TagIdx** [15] is often used to reason about fine-grained cache-related leaks.
- **Register file compression (RFC):** models the leaks induced by the register file compression optimization described in Section II.
- **Silent Store (SilStore):** differs from **RFC** in that it produces a leak when a value 0 is written in memory (instead of in a register like **RFC**). **SilStore** [5] models leaks from store operations that do not alter memory contents.
- **Multiplication simplification (mul):** models timing leaks associated with computation simplification over multiplications [5]. **mul** exposes whenever a multiplication's operands are 0 or 1.

C. RQ1: How Good is MALCOS at Synthesizing Program-Level Contracts?

RQ1 focuses on evaluating the program-level contracts generated by MALCOS during the first phase of Algorithm 1.

Experimental setup: We use MALCOS-X86 and MALCOS-ARM to synthesize program-level contracts while using the ground-truth models from Section VI-B.

For MALCOS-X86, for each target ground-truth contract, we randomly generate 100 programs and 100 random states per program, and we use MALCOS-X86 to synthesize the corresponding program-level contracts (using 100 positive examples during synthesis). We then measure the execution time and the precision/soundness of each synthesized contract against a validation set consisting of the same programs and 100 new random states per program and report the average of both metrics.

For MALCOS-ARM, for each target ground-truth contract, we randomly generate 100 programs, and we use MALCOS-ARM to synthesize the corresponding program-level contracts, with 20 counterexamples generated at each synthesis iteration and 40 positive examples overall. We then measure the execution time and the precision/soundness of each synthesized contract against a validation set consisting of the same programs and 100 new random states per program and report the average of both metrics.

Contract	MALCOS-X86		MALCOS-ARM	
	avg. P/S	avg. Time	avg. P/S	avg. Time
ct	1.0/1.0	7h 23min	1.0/1.0	13h 43m
TagIdx	0.73/1.0	12h 6min	1.0/1.0	33h 51m
RFC	0.93/1.0	10h 26min	1.0/1.0	37h 8m
mul	0.95/1.0	12h 40min	0.99/1.0	46h 53m
SilStore	0.97/1.0	3h 15min	1.0/1.0	19h 21m

TABLE II: Arch.-level results: Precision (P) & Soundness (S)

Results: Table I reports the results of our experiment. Both toolchains can successfully learn contracts that explain the target leaks with high precision (always higher than 90%), while not missing leaks (soundness at 100% for both cases). Both tools, however, perform better with simpler contracts like **ct** than with *conditional* contracts (i.e., those that produce a leakage observation under specific operand-dependent conditions) like **rfc**, **mul**, and **SilStore**, or those that produce observations exposing only part of the address like **TagIdx**. As we show in Section VI-D, some of these over-approximations are solved by learning architecture-level contracts, since the contract-minimizer takes into account multiple different programs, removing the clauses that are only necessary in one single program.

D. RQ2: How Good is MALCOS *at Learning Architecture-Level Contracts?*

RQ2 focuses on evaluating the architecture-level contracts generated by MALCOS during the second phase of Algorithm 1.

Experimental setup: We use MALCOS-X86 and MALCOS-ARM to synthesize architecture-level contracts while using the ground-truth models from Section VI-B.

For MALCOS-X86, for each target ground-truth contract, we randomly generate 1000 programs and 100 random states per program, and we use MALCOS-X86 to synthesize the corresponding architecture-level contracts (using 100 positive examples during synthesis). We then measure the overall execution time as well as the precision/soundness of the learned contracts against a validation set consisting of 10000 new random programs and 100 new random states per program, and report the average of both metrics.

For MALCOS-ARM, we generalize the program-level candidate contracts learned in Section VI-C to the corresponding architecture-level version. We then measure the execution time and the precision/soundness of each synthesized contract against a validation set consisting of 100 new programs generated by SCAM-V, and report the average of both metrics.

Results: Table II outlines the results of our experiment. Both toolchains can successfully learn contracts that explain the target leaks without missing any leaks (soundness is 100% for both cases).

MALCOS-ARM fixed the over-approximations present in the program-level contracts by generalizing them into a single contract, achieving a precision of 100%[2]. MALCOS-X86 can learn contract with high precision for all the target contracts except for **TagIdx** (73%), where the lower precision is due to MALCOS-X86 exposing more address bits than actually needed in final contract.

[2]For the **mul** contract case we conducted an additional experiment with 60 *pexs*, allowing us to achieve precision of 100%

We manually inspected the synthesized contracts (by MALCOS-X86 and MALCOS-ARM) to assess their quality. For all cases with 100% precision, we confirmed their equivalence with the ground truth. In other cases, we observed conservative over-approximations that justify not reaching 100% precision in the cases of **TagIdx**, **rfc**, **mul**, and **SilStore**.

VII. RELATED WORK

Over the years, various solutions have been proposed to analyze and mitigate hardware side-channels. We survey works related to leakage contracts and existing techniques to extract them. We also briefly review works on automating side channel analysis.

a) Leakage Contracts: Several works try to bridge the abstraction gap between hardware and software by formalizing microarchitectural features. For instance, [16], [17], [18] introduce novel operational semantics to model speculative and out-of-order execution. Some of these semantics capture leaks and speculation at high-level [16], [17], whereas others more closely resemble the microarchitecture of out-of-order processors with multiple pipeline stages, caches, and branch predictors [19], [18], [3]. Finally, other approaches, inspired by work on weak memory models, formalize microarchitectural side-effects using axiomatic semantics [20], [21], [22].

Our approach builds on the leakage contract framework [3], which precisely characterizes when a contract captures all leaks in the underlying CPU, thereby formally closing the gap between formal models and actual leaks.

b) Synthesis of Contracts: The closest work to MALCOS are [10], [9], [8]. RTL2MμPATH [10] synthesizes microarchitecture execution paths and verifies leakage contracts directly from RTL to bridge the gap between abstract models and RTL, whereas [9] construct contracts from a set of manually-specified clauses based on the results of test cases on RTL designs. In contrast, MALCOS generates accurate leakage contracts for real-world, black-box CPUs using counterexample-guided synthesis based on empirical leakage behaviors, requiring no manual templates or test-case strategies like [9]. Finally, CONJUNCT [8] synthesizes coarse-grained contracts that classify ISA instructions into "safe" or "unsafe" against timing attacks based on an RTL design, whereas MALCOS can synthesize precise and fine-grained clauses from black-box CPUs.

Beyond these, there are other relevant synthesis-based works. For instance, Liu et al. [23] uses a combination of counterexample and synteax-guided synthesis solvers to automatically synthesize a language semantics, whereas Heule et al. [24] develop a way of synthesizing the x86 ISA semantics from tests. In contrast, SYNTHCT [25] addresses cryptographic side-channel risks by synthesizing constant-time code, and PipeSynth [26] uses formal synthesis to generate axioms for microarchitectural memory consistency.

c) Automatic Reasoning about Side Channels: Apart from fuzzers used in the implementation of MALCOS, i.e., REVIZOR [6] and SCAM-V [7], [27], there exists a long list of frameworks that automate reasoning about side-channel attack, e.g., [28], [29], [30]. Osiris [28] is a fuzzing-based framework that automatically identifies timing-based side channels using a machine-readable CPU ISA specification.

CheckMate [29] offers an automated framework for synthesizing hardware exploits and crafting security tests. It lays the foundation for preemptively addressing the risks of speculative execution and side-channel attacks during the hardware design phase. Xenon [30] is a framework for formally verifying that hardware designs implemented in Verilog, are free of timing-based side-channel attacks by automatically identifying secrecy assumptions.

VIII. Conclusions

A leakage contract that describes microarchitectural leaks at ISA level is essential to enable secure programming. This paper presented a work-in-progress methodology to extract such contracts from black-box CPUs. The MALCOS synthesis tool implements this methodology for the `x86` and `ARM` architectures. Our results show that MALCOS synthesizes precise and sound contracts that capture the target leaks.

We see two concrete next steps as future work: (1) using MALCOS to perform an in-depth security analysis of x86 and ARM CPUs, and (2) extending our approach to support more complex leakage contracts, e.g., those supporting transient leaks.

References

[1] D. A. Osvik, A. Shamir, and E. Tromer, "Cache attacks and countermeasures: The case of AES," in *Proceedings of the 2006 The Cryptographers' Track at the RSA Conference on Topics in Cryptology*, ser. CT-RSA'06. Berlin, Heidelberg: Springer-Verlag, 2006, p. 1–20. [Online]. Available: https://doi.org/10.1007/11605805_1

[2] S. Briongos, P. Malagon, J. M. Moya, and T. Eisenbarth, "Reload+refresh: Abusing cache replacement policies to perform stealthy cache attacks," in *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 1967–1984. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/briongos

[3] M. Guarnieri, B. Köpf, J. Reineke, and P. Vila, "Hardware-software contracts for secure speculation," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 1868–1883.

[4] J. Hofmann, E. Vannacci, C. Fournet, B. Köpf, and O. Oleksenko, "Speculation at fault: Modeling and testing microarchitectural leakage of CPU exceptions," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 7143–7160.

[5] J. R. S. Vicarte, P. Shome, N. Nayak, C. Trippel, A. Morrison, D. Kohlbrenner, and C. W. Fletcher, "Opening pandora's box: A systematic study of new ways microarchitecture can leak private data," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 347–360.

[6] O. Oleksenko, C. Fetzer, B. Köpf, and M. Silberstein, "Revizor: Testing black-box cpus against speculation contracts," in *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2022, pp. 226–239.

[7] H. Nemati, P. Buiras, A. Lindner, R. Guanciale, and S. Jacobs, "Validation of abstract side-channel models for computer architectures," in *Computer Aided Verification - 32nd International Conference, CAV 2020 Los Angeles, CA, USA, July 21-24*, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-53288-8_12

[8] S. Dinesh, M. Parthasarathy, and C. W. Fletcher, "Conjunct: Learning inductive invariants to prove unbounded instruction safety against microarchitectural timing attacks," in *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*. IEEE, 2024, pp. 3735–3753. [Online]. Available: https://doi.org/10.1109/SP54263.2024.00180

[9] G. Mohr, M. Guarnieri, and J. Reineke, "Synthesizing hardware-software leakage contracts for RISC-V open-source processors," in *Design, Automation & Test in Europe Conference & Exhibition, DATE 2024, Valencia, Spain, March 25-27, 2024*. IEEE, 2024, pp. 1–6. [Online]. Available: https://doi.org/10.23919/DATE58400.2024.10546681

[10] Y. Hsiao, N. Nikoleris, A. Khyzha, D. P. Mulligan, G. Petri, C. W. Fletcher, and C. Trippel, "μpath: Multi-μpath synthesis with applications to hardware security verification," *arXiv preprint arXiv:2409.19478*, 2024.

[11] E. Torlak and R. Bodik, "Growing solver-aided languages with rosette," in *Proceedings of the 2013 ACM international symposium on New ideas, new paradigms, and reflections on programming & software*, 2013, pp. 135–152.

[12] G. Barthe, M. Böhme, S. Cauligi, C. Chuengsatiansup, D. Genkin, M. Guarnieri, D. M. Romero, P. Schwabe, D. Wu, and Y. Yarom, "Testing side-channel security of cryptographic implementations against future microarchitectures," in *Proceedings of the 31st ACM Conference on Computer and Communications Security*, ser. CCS 2024. ACM, 2024.

[13] L. De Moura and N. Bjørner, "Z3: An efficient smt solver," in *International conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2008, pp. 337–340.

[14] J. B. Almeida, M. Barbosa, G. Barthe, F. Dupressoir, and M. Emmi, "Verifying {Constant-Time} implementations," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 53–70.

[15] G. Doychev, B. Köpf, L. Mauborgne, and J. Reineke, "Cacheaudit: A tool for the static analysis of cache side channels," *ACM Transactions on information and system security (TISSEC)*, vol. 18, no. 1, pp. 1–32, 2015.

[16] M. Guarnieri, B. Köpf, J. F. Morales, J. Reineke, and A. Sánchez, "Spectector: Principled detection of speculative information flows," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 1–19.

[17] M. Patrignani and M. Guarnieri, "Exorcising spectres with secure compilers," in *Proceedings of the 28th ACM Conference on Computer and Communications Security*, ser. CCS 2021. ACM, 2021.

[18] R. Guanciale, M. Balliu, and M. Dam, "Inspectre: Breaking and fixing microarchitectural vulnerabilities by formal analysis," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 1853–1869. [Online]. Available: https://doi.org/10.1145/3372297.3417246

[19] R. Mcilroy, J. Sevcik, T. Tebbi, B. L. Titzer, and T. Verwaest, "Spectre is here to stay: An analysis of side-channels and speculative execution," *arXiv preprint arXiv:1902.05178*, 2019.

[20] N. Mosier, H. Lachnitt, H. Nemati, and C. Trippel, "Axiomatic hardware-software contracts for security," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, 2022, pp. 72–86.

[21] C. Disselkoen, R. Jagadeesan, A. Jeffrey, and J. Riely, "The code that never ran: Modeling attacks on speculative evaluation," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 1238–1255.

[22] H. P. de León and J. Kinder, "Cats vs. spectre: An axiomatic approach to modeling speculative execution attacks," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 235–248. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833774

[23] J. Liu, C. Murphy, A. Grover, K. J. Johnson, T. Reps, and L. D'Antoni, "Synthesizing formal semantics from executable interpreters," *Proceedings of the ACM on Programming Languages*, vol. 8, no. OOPSLA2, pp. 362–388, 2024.

[24] S. Heule, E. Schkufza, R. Sharma, and A. Aiken, "Stratified synthesis: automatically learning the x86-64 instruction set," in *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2016, pp. 237–250.

[25] S. Dinesh, G. Garrett-Grossman, and C. W. Fletcher, "Synthct: Towards portable constant-time code," in *NDSS*, 2022.

[26] C. Norman, A. Godbole, and Y. A. Manerkar, "Pipesynth: Automated synthesis of microarchitectural axioms for memory consistency," in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, 2023, pp. 513–527.

[27] P. Buiras, H. Nemati, A. Lindner, and R. Guanciale, "Validation of side-channel models via observation refinement," in *MICRO '21: 54th Annual IEEE/ACM International Symposium on Microarchitecture, Greece, October 18-22*, 2021. [Online]. Available: https://doi.org/10.1145/3466752.3480130

[28] D. Weber, A. Ibrahim, H. Nemati, M. Schwarz, and C. Rossow, "Osiris: Automated discovery of microarchitectural side channels," *CoRR*, vol. abs/2106.03470, 2021. [Online]. Available: https://arxiv.org/abs/2106.03470

[29] C. Trippel, D. Lustig, and M. Martonosi, "Checkmate: Automated synthesis of hardware exploits and security litmus tests," in *Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO-51. IEEE Press, 2018, p. 947–960. [Online]. Available: https://doi.org/10.1109/MICRO.2018.00081

[30] K. v. Gleissenthall, R. G. Kıcı, D. Stefan, and R. Jhala, "Solver-aided constant-time hardware verification," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 429–444.

Herramienta Automatizada para Identificar Datos Personales en Exfiltraciones por *Ransomware*

J.M. Tejada-Triviño, E. Castillo-Fernández, P. García-Teodoro, J.A. Gómez-Hernández
Network Engineering & Security Group (NESG)
CITIC - Universidad de Granada
juanma@darkeye.io, {elviracastillo, pgteodor, jagomez}@ugr.es

***Resumen*—Ante la constante amenaza de robo de datos por *ransomware*, se hace preciso el desarrollo de herramientas que permitan la búsqueda automática y ágil de información personal en repositorios públicos, a fin de reducir el impacto de esta lacra sobre personas y organizaciones. Aunque existen soluciones para el análisis de datos exfiltrados de brechas de seguridad, estas no suelen ser específicas para *ransomware* ni son capaces de analizar información incluida en fuentes heterogéneas. En este trabajo introducimos IDBR ("Identificación de Datos personales en Brechas de datos de Ransomware"), una herramienta novedosa que permite el análisis automático integral de repositorios de datos exfiltrados mediante *ransomware* basado en el empleo de técnicas de IA. Además de mostrar la potencialidad de una primera implementación de la herramienta con casos reales, se apuntan mejoras que permitirán hacer de IDBR una solución de alto interés para la lucha contra el cibercrimen.**

***Index Terms*—Ransomware, Data breach, Exfiltración**

Tipo de contribución: *Investigación original*

I. Introducción

El *ransomware* sigue representando un riesgo considerable, y diversos estudios evidencian su creciente impacto para la ciberseguridad global [1]. Solo en 2024, se reportaron 5.414 ataques de *ransomware* en todo el mundo, lo que representa un aumento del 11 % con respecto al año anterior [2]. De hecho, el *ransomware* se ha convertido en un negocio cada vez más lucrativo, con estimaciones de la industria que sugieren que, en 2023, los delincuentes extorsionaron más de mil millones de dólares a sus víctimas en todo el mundo [3].

El *ransomware* generalmente, impide que los usuarios accedan a sus archivos o incluso al propio entorno informático [4]. El atacante exige un rescate y, únicamente tras recibir el pago, suministra a la víctima la información necesaria para que pueda recuperar el acceso. De lo contrario, los datos pueden perderse de manera irreversible o ser expuestos públicamente. La amenaza ha evolucionado, pasando de una extorsión directa a un esquema de triple extorsión, en el cual, además de cifrar y pedir rescate o amenazar con la publicación de la información, se toma la medida adicional de amenazar a proveedores si se obtiene información sensible sobre ellos, y se ha observado también, un cambio de objetivo, de individuos a organizaciones [5], [6]. Así, los ataques de *ransomware* generan una amplia gama de perjuicios, que abarcan daños físicos, financieros, reputacionales, psicológicos y sociales. En algunos casos, la interrupción y los elevados costos derivados de estos ataques son tan críticos que ponen en peligro la continuidad de las organizaciones afectadas, llevando a algunas a cesar operaciones de forma permanente. Además, el Reglamento General de Protección de Datos (GDPR, por sus siglas en inglés) (UE 2016/679) [7] estableció la obligación para las organizaciones de notificar a la autoridad nacional de supervisión correspondiente en caso de una violación de datos personales. Este requisito de notificación ofrece varias ventajas, como la posibilidad de planificar con antelación, implementar procesos para la detección oportuna y contención de violaciones, así como evaluar el riesgo para los afectados.

Para tratar de mitigar los problemas sobrevenidos por este esquema de triple extorsión, se propone en este estudio el desarrollo de una herramienta que automatiza la identificación de información sensible en filtraciones de datos, integrando técnicas avanzadas de procesamiento multimodal. Con este objetivo, el documento se organiza como sigue. En la Sección II se exponen los antecedentes relevantes que fundamentan el desarrollo de la presente herramienta, contextualizando su necesidad en el escenario actual de ciberamenazas. En la Sección III se describe detalladamente la herramienta aquí desarrollada, su implementación técnica, una comparación con herramientas existentes y algunas mejoras innovadoras propuestas. La Sección IV está dedicada a la presentación de una prueba de concepto que ilustra los resultados obtenidos en la detección de datos críticos, para la notificación proactiva a víctimas. Finalmente, en la Sección V se exponen las conclusiones derivadas del estudio, junto con recomendaciones para futuras investigaciones.

II. Antecedentes

Existen diversas estrategias destinadas a prevenir y reaccionar ante incidentes de seguridad, y en particular ante las filtraciones de datos [8]. Además, en consecuencia con el hecho de que el GDPR exige la notificación de las filtraciones de datos, existen varias listas y repositorios [9] destinados principalmente a ayudar a las organizaciones a enfrentar la amenaza del ciberdelito y a comprender y mitigar dichos incidentes [10]. Como se ha evidenciado, se comprometen cientos de miles de registros personales (nombres, direcciones, datos de salud, información bancaria, etc.). En particular, en enero de 2024, una filtración de datos, denominada "la madre de todas las filtraciones" [11], expuso más de 26 mil millones de registros, afectando a plataformas como Twitter, Adobe, Canva, LinkedIn y Dropbox.

Para las organizaciones o individuos afectados, pagar el rescate puede parecer la única forma de minimizar la interrupción y recuperar el control. Sin embargo, desde

una perspectiva social, el pago de rescates solo alimenta el modelo de negocio de estos grupos criminales, haciendo que el *ransomware* sea más rentable y extendido [3]. Además, los ciberdelincuentes están aprovechando tecnologías y técnicas avanzadas para mover y ocultar sus fondos ilícitos. En este punto, interrumpir el ciclo de pagos es crucial para evitar que estos fondos lleguen a los criminales y para detener el crecimiento de sus operaciones.

Más allá de lo anterior, hay que destacar la relevancia de la colaboración y el intercambio de información sobre incidentes de *ransomware* [12]. Esta colaboración se expresa a menudo en términos de herramientas destinadas a investigar incidentes de seguridad y ayudar a mitigarlos. Este es el caso de varias plataformas de inteligencia diseñadas para el análisis de brechas de datos, como se muestra en la Tabla I. Sin embargo, no existen herramientas equivalentes enfocadas específicamente en filtraciones relacionadas con *ransomware*, lo que constituye un problema importante, ya que las personas cuya información se expone también tienen el derecho de conocerlo de manera oportuna.

Con el objetivo de reducir la brecha actual en las herramientas de ciberseguridad para proporcionar alertas rápidas y automatizadas de las brechas publicadas por grupos de *ransomware*, presentamos en la sección siguiente una novedosa herramienta automatizada denominada *IDBR* (”Identificación de Datos personales en Brechas de datos de *Ransomware*”), que aprovecha herramientas públicas basadas en inteligencia artificial para analizar la diversa información que se encuentra típicamente en los sitios dedicados a filtraciones de datos relacionados con *ransomware*. En particular, se consideran diversas herramientas para el reconocimiento óptico de caracteres (OCR), con el fin de confirmar la existencia de información de identificación personal (PII en adelante) en estas fugas de datos.

III. IDBR: Herramienta de Identificación de Datos Personales en Brechas de Datos de Ransomware

La creación de una herramienta automatizada para rastrear, descargar y analizar datos exfiltrados como consecuencia de ataques por *ransomware* posee una relevancia estratégica significativa en el ámbito de la ciberseguridad contemporánea. En primer lugar, la detección proactiva de fugas de datos adquiere prioridad ante el predominio de técnicas de doble o triple extorsión, utilizadas en aproximadamente el 80 % de los ataques de *ransomware* modernos [13], donde los cibercriminales no solo cifran sistemas, sino que también publicitan la información y la ofrecen al mejor postor en caso de que la víctima no pague el rescate, en foros de la *dark web* o sitios *.onion*, cuando las víctimas se niegan a pagar rescates. Disponer de una herramienta especializada permitiría identificar estas fugas en etapas tempranas, antes de su diseminación masiva, facilitando la activación oportuna de protocolos de mitigación, como la notificación a clientes afectados o la invalidación de credenciales comprometidas, reduciendo así el impacto operativo y reputacional.

La protección de datos personales representa otro eje crítico. Los documentos exfiltrados suelen contener información sensible, como números de seguridad social, historiales médicos o detalles financieros, cuya exposición genera riesgos legales y económicos. La herramienta automatizada en desarrollo que aquí se presenta, emplea técnicas avanzadas como el uso de Modelos de Lenguaje Multimodales (LMM, por sus siglas en inglés) que integran comprensión visual y lingüística, para mejorar la capacidad del reconocimiento óptico de caracteres (OCR) tradicional para extraer datos de un archivo en formato PDF [14], y la utilización de expresiones regulares, para extraer patrones específicos de PII, incluyendo números de tarjetas de crédito o direcciones físicas, con la que se podrían generar informes estructurados que faciliten el cumplimiento de normativas como la GDPR (en Europa) o la Ley Orgánica 3/2018 de Protección de Datos Personales y de garantía de derechos digitales (LOPDGDD) [15] en España. Hay que recordar que, en el caso de España, se exige un plazo de 72 horas desde que la organización tiene constancia de la brecha para notificar a la Agencia Española de Protección de Datos (AEPD) [16]. Además, esa notificación debe ser ágil ante brechas de seguridad para mitigar el impacto del incidente de seguridad.

Por otro lado, en el ámbito de la inteligencia contra amenazas, la recolección sistemática de datos de múltiples fugas permitiría identificar tendencias en las tácticas de grupos *ransomware*, como la preferencia por ciertos tipos de documentos o la explotación de vulnerabilidades recurrentes, como servidores de escritorio remoto (RDP por sus siglas en inglés) mal configurados. Además, la comparación de firmas digitales o patrones de comportamiento con aquellos asociados a grupos conocidos, como *Ransomhouse*, *Lynx*, *Fog* o *Akira*, contribuiría a la atribución de ataques y al diseño de estrategias defensivas proactivas [17].

Finalmente, en materia de respuesta forense, los metadatos incrustados en documentos filtrados, como marcas de tiempo, geolocalización o usuarios asociados a archivos, ofrecen evidencia valiosa para rastrear el origen de los ataques o identificar puntos de compromiso interno. Este tipo de información no solo respalda investigaciones legales contra actores maliciosos, sino que también fortalece la capacidad de las organizaciones para reconstruir incidentes y reforzar sus posturas de seguridad, cerrando brechas explotadas durante los ataques. La integración de estas capacidades en una única herramienta representaría, por tanto, un avance sustancial en la gestión holística de riesgos asociados al *ransomware*.

III-A. Funcionalidad de la herramienta IDBR

La herramienta que se propone en este estudio se encuentra aún en desarrollo. Sin embargo, en su fase temprana y funcional, integra un conjunto de funcionalidades novedosas críticas diseñadas para optimizar la identificación, gestión y análisis de datos exfiltrados en incidentes de *ransomware*. En primer lugar, como se muestra en la Fig. 1, el sistema incorpora un rastreador especializado *crawler*

Recurso	Descripción	Acceso
IntelX	Plataforma de inteligencia de datos y búsqueda que permite a los usuarios explorar una amplia variedad de fuentes, incluyendo filtraciones de datos, registros *WHOIS*, documentos y más.	https://intelx.io/
DeHashed	Motor de búsqueda especializado en bases de datos comprometidas.	https://dehashed.com
Have I Been Pwned?	Servicio gratuito que permite verificar si una dirección de correo electrónico o número de teléfono han sido expuestos en una filtración de datos conocida.	https://haveibeenpwned.com
Ransomwhere	Plataforma que rastrea y recruza datos sobre pagos de rescate relacionados con ataques de *ransomware*.	https://ransomwhe.re
GhostProject	Base de datos en línea que permite buscar contraseñas comprometidas y datos mediante direcciones de correo electrónico o nombres de usuario.	https://ghostproject.fr.atlaq.com
Leak Lookup	Servicio que proporciona acceso a múltiples bases de datos filtradas para verificar si la información personal ha sido comprometida.	https://leak-lookup.com/
Scylla	Motor de búsqueda que permite a los usuarios buscar en múltiples bases de datos filtradas de forma simultánea.	https://github.com/scylladb/scylladb/blob/master/dist/common/dep/scylla-env-1.0/profile.d/scylla.sh
AmILeaked	Similar a *"Have I Been Pwned"*. Servicio que permite verificar si un correo electrónico o contraseña ha aparecido en una filtración de datos conocida.	https://www.amileaked.com/
Recon-ng	Marco de trabajo de reconocimiento web que incluye módulos para buscar información en bases de datos públicas y servicios de terceros.	https://www.osintux.org/documentacion/recon-ng

Tabla I
PLATAFORMAS DE INTELIGENCIA PARA LA BÚSQUEDA DE FILTRACIONES DE DATOS.

[18], el cual es capaz de navegar de manera automatizada a través de plataformas que albergan fugas de datos, como foros en la *dark web* o sitios *.onion* (véase Tabla I), empleando tecnologías de anonimización como Tor o *proxies* distribuidos. Este componente se complementa con la capacidad de integrarse con fuentes de inteligencia externas, tales como *DarkOwl* [19] o *Recorded Future*, lo que permite enriquecer el rastreo con indicadores de compromiso (IoC en adelante) actualizados y contextualizar las amenazas identificadas.

Como complemento esencial, el motor de descarga selectiva opera bajo criterios predefinidos para filtrar documentos relevantes, priorizando formatos comunes en fugas de datos (.PDF, .docx, .xlsx, .csv, .db) o mediante la detección de palabras clave asociadas a PII, como "currículum", "contratos", "seguros", "salud" o "nóminas". Este enfoque no solo optimiza el uso de recursos computacionales, sino que también reduce el ruido en conjuntos de datos masivos, facilitando el análisis posterior.

El núcleo analítico de la herramienta reside en su detector de datos sensibles, que combina técnicas de expresiones regulares para identificar patrones estructurados, como números de seguridad social o cuentas bancarias, con modelos de *machine learning* [20] entrenados para reconocer información contextual en formatos no estructurados, tales como historiales médicos o comunicaciones internas. Esta dualidad metodológica asegura una cobertura amplia, abarcando tanto datos explícitos como aquellos que requieren interpretación semántica.

Para garantizar una respuesta ágil, se pretende incorporar un generador de alertas para que emita notificaciones en tiempo real a organizaciones cuyos dominios o activos digitales aparecen en documentos filtrados, activando protocolos de mitigación inmediatos. Adicionalmente, su interoperabilidad con plataformas SIEM (Splunk, Elasticsearch) permitiría incorporar estas alertas en flujos de trabajo existentes, potenciando la correlación de eventos y la priorización de incidentes.

Finalmente, se propone desarrollar un panel analítico unificado que ofrezca visualizaciones interactivas de patrones emergentes, como sectores económicos más afectados o tipologías de datos frecuentemente robados, apoyando la toma de decisiones estratégicas. Este módulo incluiría funcionalidades de exportación de informes detallados, adaptados a requisitos legales o regulatorios, lo que aseguraría el cumplimiento de obligaciones de transparencia y facilitaría la colaboración con equipos forenses o autoridades competentes. La sinergia entre estos componentes no solo mejora la capacidad de respuesta ante fugas de datos, sino que también establece un marco proactivo para la gestión de riesgos en un panorama de ciberamenazas en constante evolución.

III-B. Implementación

La herramienta IDBR descrita, desarrollada íntegramente en Python, se articula en módulos funcionales interconectados que abarcan todo el ciclo de procesamiento de *leaks* de *ransomware*. Como ya se ha mencionado, puede verse su estructura e interconexión entre los módulos en la Fig. 1. El primero de ellos es el *crawler*, núcleo inicial del sistema, encargado de rastrear, identificar y descargar automáticamente las filtraciones desde fuentes especializadas. Posteriormente, el módulo de clasificación gestiona la descompresión recursiva de archivos, aplica filtrados progresivos para optimizar el rendimiento, y organiza los documentos en categorías (texto, imágenes, bases de datos) para su análisis. El módulo de análisis procesa estos datos mediante técnicas de extracción de información sensible, mientras que la base de datos almacena de manera estructurada los metadatos y contenido relevante. Finalmente, el módulo de generación de informes sintetiza los hallazgos en alertas y reportes. A

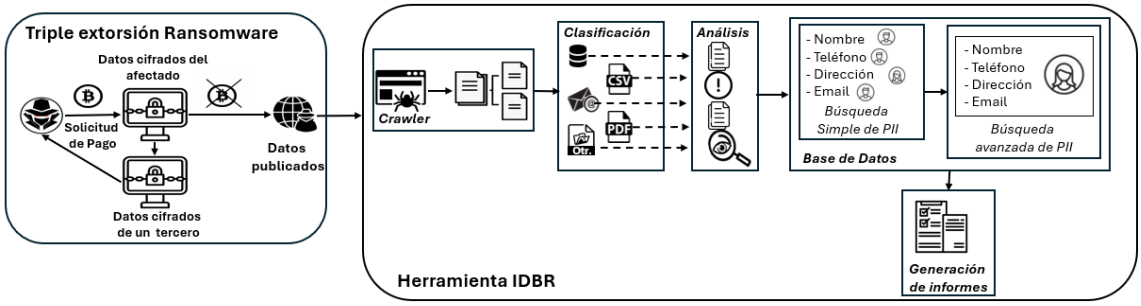


Figura 1. Funcionalidad deseada para la herramienta IDBR.

continuación, se detalla el diseño, alcance y funcionamiento técnico de cada componente:

- El *módulo crawler* se vale de rastreadores de código abierto como *Ransomwatch, Breach house, Ransomware.live* y *Ransomlook.io* para monitorizar y recuperar enlaces a filtraciones de datos tan pronto como estas se hacen públicas. Una vez identificadas, las filtraciones se descargan mediante un proceso estructurado que combina un *script* basado en *Aria2-Onion-Downloader* [21] para facilitar la transferencia eficiente de archivos a través de la red Tor. Para tratar de superar las limitaciones de ancho de banda inherentes a Tor, se utilizan:
  1. *Torget* [22], que es una herramienta de línea de comandos que emplea conexiones paralelas para optimizar la velocidad y eficiencia de las descargas.
  2. *Torboost* [23], que acelera las transferencias mediante la optimización de conexiones y el uso de múltiples hilos.
- Una vez completada la descarga integral de la filtración, entra en acción el *módulo de clasificación*, donde se inicia el proceso de descompresión recursiva de todos los archivos comprimidos (en formatos .Zip, .7z, .targz o .RAR), garantizando que ningún archivo quede sin extraer. Este paso es fundamental para acceder al contenido bruto y asegurar la integridad del conjunto de datos antes de su análisis. Posteriormente, se implementa un filtrado progresivo y estratificado, estrategia diseñada para optimizar la eficiencia operativa y minimizar errores. En esta fase, los archivos se clasifican en categorías específicas: archivos de texto, documentos estructurados (*p.ej.*, DOCX), PDF, imágenes y bases de datos. Esta segmentación facilita el procesamiento posterior especializado de cada tipo de archivo, adaptando las técnicas analíticas a sus particularidades de formato.
  Mediante este enfoque se obtiene un conjunto integral de datos de la entidad comprometida, que incluye archivos internos, facturas, comunicaciones y documentos confidenciales. Sin embargo, el interés principal se centra en la PII, la cual, en filtraciones a gran escala, con volúmenes del orden de varios TB, suele representar una fracción mínima del total (*p.ej.*, 1 GB).

- Para quedarse con la información interesante, en el *módulo de análisis*, a los archivos de texto se les aplican técnicas de procesamiento multimodal para extraer y clasificar automáticamente la PII, priorizando su identificación dentro de conjuntos de datos masivos. Así, en esta primera versión de la herramienta, se ejecuta un algoritmo de detección de PII, centrado inicialmente en direcciones de correo electrónico. Para ello se emplea una expresión regular (*regex*) basada en la especificación RFC 2822, estándar que define la sintaxis válida para direcciones de correo electrónico. Cuando se identifica una coincidencia, el correo electrónico se convierte a formato JSON (Fig. 2), estructura que permite su indexación eficiente en bases de datos como *Elasticsearch* [24]. Adicionalmente, se captura un contexto textual de 100 caracteres precedentes y posteriores al punto de detección, proporcionando información semántica sobre el uso del correo electrónico dentro del documento.
  Tras el análisis de archivos de texto, se procede al tratamiento diferenciado de documentos PDF. Estos se categorizan en dos grupos según sus características: PDF digitalizados, que contienen capas de texto incrustado, y PDF escaneados, documentos sin texto legible, compuestos únicamente por imágenes. Los primeros se procesan mediante el mismo algoritmo aplicado a los archivos de texto. Los segundos se analizarán en próximas versiones de la herramienta.
  Los documentos identificados como portadores de información crítica, se someten a un análisis avanzado mediante Modelos de Lenguaje Multimodales (LMM). Estas variantes de Modelos lingüísticos de gran tamaño (LLM), enriquecidas con capacidades OCR [25], permiten la extracción precisa de texto incluso en condiciones adversas (resolución baja, fondos complejos). Al combinar aprendizaje profundo y razonamiento contextual, los LMM interpretan tanto el contenido textual como visual, superando las limitaciones de los sistemas OCR tradicionales.
  En el caso de archivos estructurados (CSV, XML, SQL, NoSQL, JSON), se aplica un protocolo trifásico. Primero, se inspeccionan los nombres de columnas o campos para detectar términos indicativos como ”email” o ”nombre”, extrayendo automáticamente los datos correspondientes. Si los encabezados no son explícitos, se escanean las primeras 30 líneas del archivo utilizando *regex* para identificar patrones asociados a PII. Finalmente, los datos filtrados se transforman a JSON estandarizado. Para garantizar la unicidad de los registros en la base de datos, se genera

un *hash* SHA-1 de cada entrada. Este mecanismo evita duplicados durante la indexación en la base de datos, asegurando la consistencia del repositorio y optimizando las consultas posteriores. Este enfoque metodológico no solo prioriza la precisión en la identificación de PII, sino que también establece un marco reproducible para el análisis de grandes volúmenes de datos en entornos de ciberseguridad.
Esta metodología integral, que abarca desde el procesamiento de documentos hasta el análisis de bases de datos heterogéneas, establece un marco reproducible para la identificación eficiente de información sensible en entornos de *ransomware*. La combinación de técnicas clásicas (*regex*) con modelos de IA multimodal (LMM) garantiza una cobertura amplia y adaptable a la evolución de las tácticas de extorsión digital. Además, este método de análisis y procesamiento permite localizar de manera proactiva a los afectados.

- La *base de datos* constituye el núcleo central del sistema propuesto, diseñada para almacenar y estructurar toda la información obtenida del módulo de análisis mediante un proceso de indexación avanzada. Al tratarse de una base de datos *no relacional (NoSQL)*, emplea un modelo basado en colecciones y documentos, lo que permite una gestión flexible de datos heterogéneos y no estructurados, como archivos de texto, metadatos o registros de actividad. Esta arquitectura facilita la escalabilidad horizontal y el almacenamiento distribuido, críticos para manejar volúmenes masivos de información derivados de fugas de datos.
  Además de permitir búsquedas eficientes y la correlación de información proveniente de múltiples fuentes asociadas a una misma filtración, la base de datos incorpora funcionalidades avanzadas para la identificación y clasificación de PII. Mediante consultas personalizadas y agregaciones, es posible rastrear patrones recurrentes, como la aparición de direcciones de correo electrónico o números de identificación fiscal en documentos dispersos, estableciendo relaciones entre entidades afectadas y contextos de compromiso.
  La base de datos también habilita el procesamiento de lenguaje natural (NLP) para detectar información sensible en texto no estructurado, como descripciones clínicas en historiales médicos, complementando los métodos basados en expresiones regulares.
  Finalmente, la capacidad de generar agregaciones estadísticas y modelos gráficos de relaciones entre entidades (mediante grafos) proporciona una base sólida para la investigación forense, el análisis de tendencias y la generación de informes técnicos adaptados a requerimientos legales o regulatorios. Esta multifuncionalidad posiciona a la base de datos no solo como un repositorio, sino como una plataforma analítica integral para la gestión proactiva de riesgos en entornos de ciberseguridad dinámicos.

- Por último, el *módulo de generación de informes* realiza consultas a la base de datos para generar los informes en base a la información requerida. En este contexto, la herramienta que se presenta desempeña un papel fundamental, ya que extrae e indexa la PII contenida en las filtraciones de *ransomware*, permitiendo realizar búsquedas eficientes y la posibilidad de alertar a los afectados. Mediante un análisis sistemático de los conjuntos de datos filtrados, la herramienta garantiza que las víctimas puedan ser notificadas de manera oportuna, proporcionándoles la posibilidad de implementar acciones que mitiguen los riesgos derivados de la exposición de sus datos.

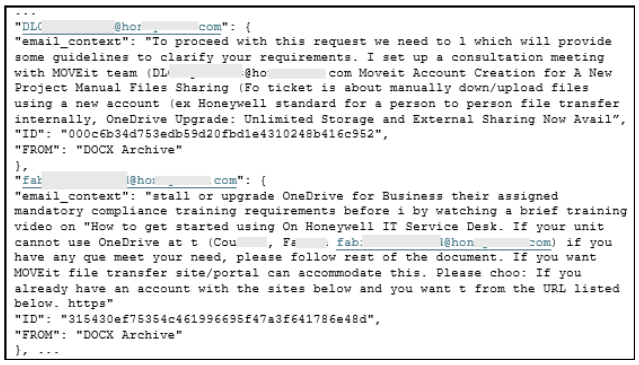


Figura 2. Ejemplo de búsqueda de email entre documentos exfiltrados.

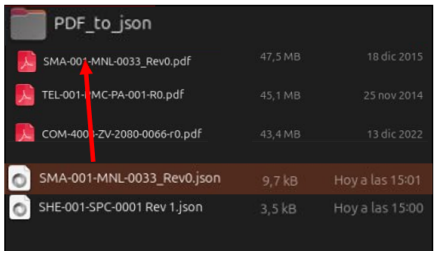


Figura 3. Ejemplo de extracción de PDF, búsqueda PII y conversión a JSON.

III-C. Comparativa con otras herramientas existentes

La creciente sofisticación de los ataques de *ransomware* ha impulsado el desarrollo de herramientas especializadas en la monitorización y análisis de filtraciones de datos. Como se esboza en la Tabla II, que resume las capacidades clave de distintas soluciones, plataformas como *Recorded Future* destacan por su enfoque en inteligencia de amenazas predictiva, integrando datos de fuentes abiertas (OSINT) y *dark web*. Sin embargo, su capacidad para procesar contenido específico de fugas de datos, como metadatos o texto embebido en imágenes, es limitada, priorizando en cambio la correlación de indicadores de compromiso (IoC) a gran escala. Por su parte, *DarkOwl* ofrece un acceso profundo a foros clandestinos y bases de datos filtradas, pero carece de mecanismos automatizados para identificar información sensible (*p.ej.*, PII dentro de los archivos exfiltrados), lo que reduce su utilidad en contextos de respuesta inmediata.

Herramientas como *Flare.io* se centran en la detección proactiva de infraestructuras expuestas y conversaciones de cibercriminales en la *dark web*, aunque su alcance



Figura 4. Ejemplo de búsqueda de PII en la base de datos.

no incluye el análisis forense de *leaks* publicados en sitios de fuga de datos consecuencia de *ransomware*. En contraste, plataformas agregadoras como *RansomLook.io* y *Ransomware.live* compilan listados actualizados de víctimas y enlaces a filtraciones, pero operan como repositorios estáticos: no analizan el contenido de los datos expuestos ni generan alertas dirigidas a afectados. Esta desconexión entre la recolección y el procesamiento profundo de *leaks* representa una brecha crítica en el ciclo de gestión de amenazas.

Nuestra herramienta IDBR aborda estas limitaciones mediante un enfoque multimodal que combina técnicas avanzadas de procesamiento de datos. A diferencia de otras alternativas, no solo descarga *leaks* de sitios especializados en *ransomware*, sino que realiza un análisis forense automatizado que incluye:

- Extracción de texto y metadatos (autor, fechas de modificación) de documentos estructurados (PDF, hojas de cálculo).
- Reconocimiento de texto en imágenes mediante OCR.
- Correlación de PII con bases de datos públicas.

Esta integración de capacidades técnicamente diversas, ausente en herramientas como *DarkOwl* o *Ransomware.live*, permite una respuesta más ágil y precisa frente a filtraciones masivas, donde el volumen de datos (hasta 1 TB por incidente) dificulta la detección manual de información crítica.

Esta capacidad combinada entre la automatización en la descarga y clasificación de *leaks*, posiciona a la IDBR como una solución integral frente a herramientas fragmentadas como *Flare.io* (enfocada en *dark web*) o *RansomLook.io* (limitada a agregación de enlaces).

Característica	Recorded-Future	Dark-Owl	Flare.io	Ransom-Look.io	IDBR
Fuentes de datos	OSINT + dark web	Dark web	Dark web	Ransom-ware leaks	Ransom-ware leaks + dark web
Análisis de texto en documentos	Sí	Sí	Limitado	No	Sí (avan-zado)
Procesamiento de metadatos	No	No	No	No	Sí
Reconocimiento de imágenes (OCR+IA)	No	No	No	No	Sí (multi-modal)
Alertas a víctimas	No	No	No	No	Sí
Integración con crawlers	Sí (APIs)	Sí	Sí	Sí (leaks)	Sí (automática)
Enfoque en *ransomware*	Secundario	No	No	Principal	Principal + análisis forense

Tabla II
COMPARATIVA CON HERRAMIENTAS SIMILARES EXISTENTES

IV. CASOS DE USO

Para validar las capacidades operativas de la herramienta IDBR, se han analizado cuatro incidentes paradigmáticos de exfiltración de datos derivados de ataques de *ransomware*, cuyos resultados se sintetizan en la Tabla III. En ella se esbozan algunos de los hallazgos que se pueden deducir del análisis realizado con IDBR a las corporaciones afectadas por *ransomware* objeto de análisis en los siguientes casos de uso.

IV-A. Caso 1: Pension Benefit Information (pbInfo)

El grupo de *ransomware Cl0p*, vinculado a operadores rusos, perpetró un ciberataque contra *Pension Benefit Information (pbInfo)* [26], proveedor estadounidense especializado en soluciones de gestión de datos poblacionales, exfiltrando 33,21 GB de información confidencial, incluyendo datos de beneficiarios de pensiones y seguros. La herramienta desarrollada por los autores, IDBR, ha identificado que 8,9 GB corresponden a PII, distribuida en 783 bases de datos, 702 archivos PDF y 72 documentos de texto. El análisis automatizado revela la exposición de 79.554 nombres completos, 88.674 correos internos y 204 documentos identificativos (DNI/pasaportes), lo que incrementa el riesgo de suplantación de identidad y fraudes financieros.

IV-B. Caso 2: Pioneer Electronics

El mismo grupo *Cl0p* atacó a *Pioneer Electronics* [27], empresa japonesa, exfiltrando 114,14 GB de datos. Como se ilustra en la Fig. 3 que muestra la metodología de procesamiento de IDBR, los archivos PDF identificados se someten a un flujo de conversión automatizada. La mencionada figura esboza que la carpeta *PDF_to_json* alberga documentos originales descargados de la filtración durante el análisis y procesado de la información contenida en los PDF, en la misma carpeta se añaden los archivos *.json* homónimos, generados tras detectar PII en los PDF. Cada JSON estructura la información sensible extraída (nombres, correos, identificadores) en índices estructurados, preservando una relación nominal unívoca con el PDF fuente (*p.ej.*, SMA001-MNL003_Rev0.pdf → SMA001-MNL003_Rev0.json). Este enfoque no solo agiliza la indexación en la base de datos no relacional de IDBR,

Exfiltración de datos tras ataque *ransomware*		Tamaño total de datos		Nº total de archivos encontrados y analizados por tipo de archivo que contienen PII					Nº total de PII por tipo de dato econtrados en los documentos analizados			
Empresa	**País**	**Datos exfiltrados**	**PII**	**.txt**	**PDF**	**.doc**	**Imágenes**	**Bases de Datos**	**Nº nombres**	**Nº correos internos**	**Nº correos externos**	**Nº DNI/ Pasaportes**
pbInfo	EEUU	33,21 GB	8,9 GB	72	702	36	23	783	79.554	88.674	2.695	204
Pioneer Electronics	Japón	114,14 GB	24,6 GB	233	1.361	84	106	1.352	195.443	104.875	95.471	582
Honeywell	EEUU	233,45 GB	22,2 GB	159	1.714	13	145	2.542	115.961	165.343	24.893	373
Viasat	España	97,66 GB	17,3 GB	98	1.136	141	63	3.173	138.749	38.715	88.451	327

Tabla III
HALLAZGOS DEL ANÁLISIS REALIZADO CON IDBR A CUATRO ORGANIZACIONES AFECTADAS POR RANSOMWARE.

sino que también garantiza la trazabilidad durante auditorías forenses o análisis de impacto.

La eficacia de este proceso se refleja en la correlación precisa entre los 1.361 PDF analizados y los 1.352 JSON generados, donde el 99.3 % de los archivos contenían PII explotable. Este hallazgo subraya la prevalencia de documentos administrativos como vectores críticos de filtración en ataques a empresas manufactureras como *Pioneer Electronics*.

IV-C. Caso 3: Honeywell

En el ataque a *Honeywell* [28], se exfiltraron 233,45 GB, de los cuales 22,2 GB albergaban PII. La herramienta IDBR detectó 1.714 archivos PDF predominantemente documentos administrativos con PII (nóminas, contratos laborales, registros fiscales) y 2.542 bases de datos, conteniendo 165.343 correos internos y 24.893 correos externos. Como se observa en la Fig. 4, que muestra un registro tipo de las bases de datos analizadas, la información expuesta incluye datos laborales y personales críticos:

- Identificadores únicos (*NºEmpleado TRESS*).
- Datos demográficos (*Nombre, Fecha de nacimiento, RFC (Registro Federal de Contribuyentes mexicano)*).
- Información salarial (*Salario Mensual, Salario Diario*).
- Detalles corporativos (*Email, Compañía*).

Este nivel de detalle, particularmente el RFC y los montos salariales, podría facilitar fraudes fiscales, suplantación institucional o ataques de ingeniería social personalizados. La estructura de los correos electrónicos (nombre.apellido@dominio) refuerza el riesgo de campañas de *phishing* creíbles contra empleados.

IV-D. Caso 4: Targa Viasat España

El grupo *Medusa* comprometió a *Targa Telematics* (*Targa Viasat España*) [29], empresa española de plataformas digitales para la movilidad. Exigió 138.689 euros a cambio de los datos robados y finalmente la banda filtró los 97,66 GB de datos secuestrados [30]. La herramienta IDBR identificó 17,3 GB de PII, con 3.173 bases de datos y 1.136 archivos PDF como formatos predominantes. El análisis cuantificó 138.749 nombres, 88.451 correos externos (posiblemente de clientes) y 327 documentos identificativos, lo que refleja un riesgo elevado de fraude transnacional y violaciones al GDPR e implica sanciones por la AEPD si no se comunica el incidente en 72 horas desde que ocurre.

IV-E. Implicaciones Generales

Se ha utilizado la herramienta IDBR para localizar, descargar, descomprimir, analizar y extraer los PII de los documentos, imágenes y bases de datos exfiltradas y poder generar estadísticas sobre la información encontrada. Los resultados obtenidos evidencian patrones comunes:

- Predominio de PDF y bases de datos (70 % de los archivos analizados), indicativo de la priorización de información estructurada por los atacantes.
- Correlación geográfica: Las empresas estadounidenses (*PBI* y *Honeywell*) presentan mayor volumen de correos internos (˜85.000 en promedio), mientras que la europea (*Viasat*) destaca en correos externos (˜88.000), sugiriendo diferencias en los flujos de comunicación corporativa.

Si analizamos la eficacia de IDBR, la herramienta redujo el tiempo de análisis en un 92 % frente a métodos manuales, según métricas internas, validando su utilidad en la gestión proactiva de brechas.

Estos casos subrayan la necesidad de integrar herramientas como IDBR en estrategias de ciberseguridad, especialmente ante la sofisticación creciente de grupos como *Cl0p* o *Medusa*, cuya explotación de vulnerabilidades críticas (*p.ej.*, CVE-2023-34362 [31] en MOVEit) demanda respuestas automatizadas y basadas en datos.

V. CONCLUSIONES

A pesar de la existencia de herramientas para el análisis de exfiltraciones de datos en busca de información personal sensible, se observa que todas ellas son mejorables en aspectos como no especificidad para brechas *ransomware*, uso de repositorios estáticos, no análisis de fuentes de datos heterogéneas, etc. Ante esta situación, los autores hemos desarrollado la herramienta *IDBR*, la cual trata de dar solución a las limitaciones anteriores.

No obstante los buenos resultados obtenidos en pruebas experimentales, se pretende seguir mejorando la herramienta en diversos apectos. Por una parte, la inclusión de un módulo de análisis de imágenes que permita acceder a información sensible como DNI, pasaportes o cédulas identificativas de personas de diferentes países y escritas en diferentes idiomas incluso utilizando alfabetos específicos como el chino, ruso o el griego.

Adicionalmente, la herramienta IDBR podría superar las limitaciones de los sistemas OCR tradicionales, que suelen generar errores en condiciones adversas (baja resolución, fondos complejos), mediante el empleo der Modelos de Lenguaje Multimodales (LMM) entrenados en contextos lingüísticos y visuales. De este modo, la herramienta no solo extraería texto crudo de imágenes, sino que inferiría significado semántico. Por ejemplo, diferenciando entre un número de tarjeta de crédito y una secuencia numérica aleatoria en una captura de pantalla.

Otras líneas de mejora podrían consistir en:

- Realización de análisis temporales para mapear la evolución de una filtración, identificando ventanas de tiempo críticas en las que los datos fueron accedidos o modificados.
- Integración de técnicas de geolocalización en los documentos indexados para asociar fugas de datos con regiones o infraestructuras específicas, mientras que la vinculación con *feeds* externos de inteligencia de amenazas (como indicadores de compromiso o *hashes* de archivos maliciosos) enriquecería el contexto analítico.
- Enmascaramiento de datos y cifrado de extremo a extremo, asegurando que los PII identificados no sean expuestos durante consultas o exportaciones.

Gran parte de lo anterior se encuentra ya en proceso de desarrollo por los autores y es intención de estos la oportuna publicación de la funcionalidad completa de la herramienta una vez esté plenamente operativa. Asimismo, se prevé la disposición pública de IDBR para su libre uso por parte de la comunidad.

Agradecimientos

Este trabajo ha sido desarrollado dentro del "Plan de Recuperación, Transformación y Resiliencia", convenio C025/24 INCIBE-Universidad de Granada, financiado por la Unión Europea (fondos *Next Generation*).

Referencias

[1] Cisco Umbrella, "From Trojan Takeovers to Ransomware Roulette", en *Cisco Cyber Threat Trends Report https://umbrella.cisco.com/info/cyber-threat-trends-report*, 2025.

[2] Check Point, "Ransomware Annual Report 2024", en *https://cyberint.com/blog/research/ransomware-annual-report-2024*, 2024.

[3] Home Office, "Ransomware Legislative Proposals: Reducing Payments to Cyber Criminals and Increasing Incident Reporting", en *Government Consultation https://assets.publishing.service.gov.uk/media/67864097c6428e013188175a/Consultation-Document-Proposals-v2.pdf*, 2025.

[4] N.A. Hassan, "Ransomware Revealed: A Beginner's Guide to Protecting and Recovering from Ransomware Attacks", en *Apress*, 2019 (ISBN:978-1484242544).

[5] M. Aggarwal, "Ransomware Attack: An Evolving Targeted Threat", en *14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-7, 2023 (DOI: https://doi.org/10.1109/ICCCNT56998.2023.10308249).

[6] J. MacColl, P. Husch, G. Mott, J. Sullivan, J. R.C. Nurse, S. Turner, N. Pattnaik, "The Scourge of Ransomware: Victim Insights on Harms to Individuals, Organisations and Society", en *Royal United Services Institute https://www.rusi.org/explore-our-research/publications/occasional-papers/ransomware-victim-insights-harms-individuals-organisations-and-society*, 2024.

[7] E. Union, Regulation (eu), "2016/679 on the protection of natural persons regarding data processing (GDPR)", en *https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng*, 2016.

[8] W. Fisher, R.E. Craft, M. Ekstrom, J. Sexton, J. Sweetnam, "Data Confidentiality: Detect, Respond to, and Recover from Data Breaches", en *NIST Special Publication 1800-29 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1800-29.pdf*, February 2024.

[9] C.D. Hylender, P. Langlois, A. Pinto, S. Widup, "2024 DBIR Data Breach Investigation Report", en *Verizon https://www.verizon.com/business/resources/reports/dbir*, 2024.

[10] America's Cyber Defense Agency, "StopRansomware: BianLian Ransomware Group", en *https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-136a*.

[11] V. Petkauskas, "Mother of all breaches reveals 26 billion records: what we know so far", en *Cybernews https://cybernews.com/security/billions-passwords-credentials-leaked-mother-of-all-breaches*,29 January 2024.

[12] Centre for Cybersecurity Belgium, "Uncovering Patterns between GDP Size and Ransomware Gangs' Choice for Targets", en *https://ccb.belgium.be/sites/default/files/RansomwareResearchReport_GDP%26targeting_2024-02-01.pdf*, 2024.

[13] HACQUEBORD, Feike; HILT, Stephen; SANCHO, David, "The Near and Far Future of Ransomware Business Models", en *Trend Micro Research, https://www.key4biz.it/wp-content/uploads/2022/12/wp-the-near-and-far-future-of-ransomware.pdf*, 2022.

[14] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X. Yin, C. Liu, L. Jin, X. Bai, "OCRBench: on the hidden mystery of OCR in large multimodal models.", en *Sci. China Inf. Sci.*, 67, 220102, 2024 (DOI: https://doi.org/10.1007/s11432-024-4235-6).

[15] LOPDGDD, "Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.", en *Boletín Oficial del Estado https://www.boe.es/eli/es/lo/2018/12/05/3/con*, ≪BOE≫ núm. 294, de 06 diciembre 2018.

[16] Agencia Española de Protección de Datos, "Notificación de brechas de datos personales a la Autoridad de Control", *AEPD, https://www.aepd.es/derechos-y-deberes/cumple-tus-deberes/medidas-de-cumplimiento/brechas-de-datos-personales-notificacion*, 24 de Febrero de 2025.

[17] M. Verma, P. Kumarguru, S. Brata Deb, A. Gupta, "Analysing Indicator of Compromises for Ransomware: Leveraging IOCs with Machine Learning Techniques", en *2018 IEEE International Conference on Intelligence and Security Informatics (ISI), Miami, FL, USA,*, pp. 154-159, 2018 (DOI: https://doi.org/10.1109/ISI.2018.8587409).

[18] A. Dalvi, P. Kulkarni, A. Kore, SG Bhirud, "Dark Web Crawling for Cybersecurity: Insights into Vulnerabilities and Ransomware Discussions", en *2nd International Conference for Innovation in Technology (INOCON) , Bangalore, India*, págs. 1-6, 2023 (DOI: https://doi.org/10.1109/INOCON57975.2023.10101162).

[19] A. Dalvi, S. Bhirud, "Dark web monitoring as an emerging cybersecurity strategy for businesses", en *Int. J. Inf. Eng. Electron. Bus.(IJIEEB)*, 16(2), 54-67, 2024.

[20] A. Alhogail, A. Alsabih, "Applying machine learning and natural language processing to detect phishing email.", en *Computers & Security*, 110, 102414, 2021.

[21] sn0b4ll, .ªria2-onion-downloader",*https://github.com/sn0b4ll/aria2-onion-downloader*, 2024.

[22] mtrojnar, "torget" *https://github.com/mtrojnar/torget*, 2024.

[23] tasooshi , "toboost",*https://github.com/tasooshi/torboost*, 2022.

[24] Elastic, "Elasticsearch: Real-time Search and Analytics Engine", en *https://www.elastic.co/elasticsearch*, 2024.

[25] Zhang, Yuchen and others,"Mini-Monkey: A Vision-Language Model for Document Understanding", en *arXiv preprint https://arxiv.org/abs/2408.02034* , 2408.02034, cs.CL,2024.

[26] V. Petkauskas, "Breach of death auditor PBI exposes details of 370,000 people", en *Cybernews https://cybernews.com/news/pbi-data-breach-moveit*, julio 2023.

[27] Stefanie Schappert, "TomTom, Pioneer Electronics, Autozone latest to navigate MOVEit attacks", *Cybernews, https://cybernews.com/news/tomtom-navigates-moveit-attacks-clop/*, noviembre 2023.

[28] V. Petkauskas, "Honeywell hack exposed nearly 120K people", *Cybernews https://cybernews.com/news/honeywell-breach-exposed-thousands/*, noviembre 2023.

[29] EUROPA PRESS , "Targa Viasat sufre un ciberataque que compromete casi 100 GB de documentos financieros y personales", *EUROPA PRESS, https://www.europapress.es/motor/sector-00644/noticia-targa-viasat-sufre-ciberataque-compromete-casi-100-gb-documentos-financieros-personales-20240704185418.html*, 2024.

[30] M.Luz Domínguez, "Un ciberataque a Targa Viasat compromete casi 100 GB de datos financieros y personales", *Bit Life Media, https://bitlifemedia.com/2024/07/un-ciberataque-a-targa-viasat-compromete-casi-100-gb-de-datos-financieros-y-personales/*, julio 2024.

[31] Mitre, "CVE-2023-34362", en *Mitre https://www.cve.org/CVERecord?id=CVE-2023-34362*, junio 2023.

Sistema Generador de Conjuntos de Datos con Ciberataques

Á. Sánchez-Corchado
Universidad de Cádiz
Escuela Superior de Ingeniería
angel.sancor@alum.uca.es

L. Gutiérrez-Galeano
Universidad de Cádiz
Escuela Superior de Ingeniería
leopoldo.gutierrez@uca.es

J.J. Domínguez-Jiménez
Universidad de Cádiz
Escuela Superior de Ingeniería
juanjose.dominguez@uca.es

***Resumen*—En la actualidad, la seguridad informática o ciberseguridad, se ha convertido en un pilar fundamental en el ámbito de la tecnología debido al crecimiento exponencial de ciberataques. A medida que evoluciona la tecnología, aparecen nuevas técnicas, herramientas, dispositivos, aplicaciones, etc. En malas manos, provocan que los ciberdelincuentes cuenten con herramientas más potentes y eficaces con las cuales son capaces de llevar a cabo ataques informáticos. Generando, por tanto, un problema constante para las personas.**

La necesidad de conjuntos de datos para la investigación y el desarrollo en ciberseguridad ha sido un desafío persistente durante décadas. Si bien el reciente auge de la inteligencia artificial ha intensificado la atención sobre la detección y prevención de ciberamenazas, la dependencia de datos de calidad para entrenar y evaluar modelos no es un concepto nuevo. De hecho, la disponibilidad de datasets relevantes y actualizados ha sido un factor crítico en el avance del campo desde sus inicios. Sin embargo, para la aplicación de técnicas como *machine learning* y *deep learning*, es fundamental contar con *datasets* actuales y realistas para poder desarrollar correctamente sistemas de detección de intrusiones (IDS), sistemas de prevención de intrusiones (IPS) o sistemas de gestión de eventos e información de seguridad (SIEM) entre otros. Actualmente, los conjuntos de datos públicos son escasos o desactualizados, lo que limita la eficacia de estas herramientas para la detección de ciberataques.

Este artículo tiene como objetivo desarrollar un sistema capaz de solventar esta problemática, permitiendo simular tráfico de red realista en un entorno virtual controlado. Utilizando herramientas como Docker y Docker Compose, se crearán diferentes entornos virtuales en los que distintos contenedores representarán atacantes y víctimas. Estos entornos virtualizados generarán tráfico tanto benigno como malicioso, permitiendo la captura y etiquetado de datos para su uso en sistemas de detección basados en inteligencia artificial.

***Index Terms*—Ciberataques, Detección, Virtualización, Datasets**

Tipo de contribución: *Investigación original*

I. Introducción

El crecimiento vertiginoso de Internet ha impactado en todos los ámbitos de la vida moderna. Cada vez más personas y dispositivos se mantienen interconectados; a su vez aparecen y se materializan nuevos tipos de ciberataques que ponen en riesgo la disponibilidad, confidencialidad e integridad de la información. Considerando la amplia tipología de amenazas que pueden afectar a un sistema informático, es necesario hacer uso de una variedad de técnicas y herramientas para minimizar el impacto de los constantes ataques que se producen en las redes informáticas. Aun así, esto sigue siendo insuficiente, ya que los ciberdelincuentes siempre están a la vanguardia y desarrollan continuamente nuevos tipos de ciberamenazas [1].

La intrusión, entendida como el ataque a los sistemas informáticos conectados en red con el objetivo de afectar a la disponibilidad, integridad o la confidencialidad de la información, es amplia y diversa. Esto incluye ataques con malware, denegación de servicio (DoS), phishing o ingeniería social, ataques de inyección SQL, entre otros tantos [2].

Se encuentra disponible una amplia gama de herramientas cuyo objetivo es la protección de la información. En términos de protección de la red de una organización [3], existen los denominados sistemas de gestión de eventos e información de seguridad o SIEM [4], sistemas de prevención de intrusiones o IPS [5], y sistemas de detección de intrusiones o IDS [6]. La característica principal de todas estas herramientas mencionadas anteriormente es la detección de ciberataques.

Sin embargo, las organizaciones se enfrentan continuamente a diversas problemáticas para proteger su infraestructura digital de ataques cada vez más complejos y avanzados. Es importante mencionar que si estamos haciendo referencia a una infraestructura crítica, los efectos de un ciberataque exitoso podrían ser devastadores. El incremento en cuanto a cantidad de ciberataques es notable a medida que pasa el tiempo. Según Verizon [7] el 68 % de las brechas de seguridad están vinculadas a un factor humano, ya sea causado por un ataque de ingeniería social, como pueden ser el phishing, el vishing, el pretexting, etc., o por errores que comete el usuario, como pueden ser la creación de contraseñas débiles, la no realización de actualizaciones de seguridad o la mala configuración de las redes, entre muchas otras causas.

Esto evidencia el hecho de que, incluso con soluciones y técnicas robustas, las organizaciones siguen siendo vulnerables ante cierto tipo de vulnerabilidades como, por ejemplo, aquellas dependientes del comportamiento humano. A causa de ello, las soluciones implantadas en las organizaciones podrían no ser efectivas incluso siendo adecuadas. Esto genera un impacto económico crítico en cualquier tipo de organización. IBM [8] revela que el coste promedio de una filtración de datos alcanzó los 4,88 millones de dólares en 2024, un incremento del 10 % respecto a 2023. Además, el mismo estudio destaca que las organizaciones que hacen uso de soluciones basadas en técnicas de Inteligencia Artificial reducen los costes asociados en un 24 %. En la actualidad, la investigación en el campo de la detección de ciberataques se encuentra en auge. De hecho, continuamente aparecen nuevas publicaciones afines a esta línea de investigación, la mayoría de ellas haciendo uso de técnicas basadas en Inteligencia Artificial. Luego, esta línea de investigación es viable, siempre y cuando se disponga de conjuntos de datos que contengan una amplia variedad de tipos de ataques que sean actuales.

Para ello, se espera que existan herramientas actualizadas

que puedan prevenir estos tipos de ataques informáticos que afectan en la vida de las personas. Sin embargo, lo que vemos en la actualidad es que en muchas ocasiones estas herramientas están muy desactualizadas. Un claro ejemplo de ello son los sistemas de detección de ciberataques que, en gran cantidad de ocasiones, utilizan conjuntos de datos desactualizados. A lo largo de la historia de la investigación en ciberseguridad, se han desarrollado varios conjuntos de datos con el objetivo de facilitar la evaluación de sistemas de detección de intrusiones. Ejemplos notables, aunque con sus propias limitaciones reconocidas, incluyen el dataset KDD Cup 99 (derivado de datos de 1998) y el dataset DARPA/Lincoln Laboratory de 1999. Estos conjuntos de datos pioneros sentaron las bases para la investigación posterior, pero también pusieron de manifiesto los desafíos inherentes a la creación de datasets representativos y actualizados. Sin embargo, la creciente sofisticación de las técnicas de inteligencia artificial aplicadas a la ciberseguridad introduce nuevas dimensiones a este desafío. Los modelos de aprendizaje automático, en particular las arquitecturas de aprendizaje profundo, requieren volúmenes de datos significativamente mayores y más diversos para lograr un rendimiento óptimo y una buena generalización. Además, la capacidad de la IA para identificar patrones complejos y sutiles exige que los datasets no solo contengan una amplia gama de ataques conocidos, sino también representaciones de amenazas emergentes y comportamientos anómalos que puedan pasar desapercibidos para los sistemas tradicionales.

Este artículo presenta el desarrollo de una herramienta para la generación de casos de pruebas, los cuales nos permitirán solventar problemas como la escasez de conjuntos de datos públicos enfocados en la detección de ciberataques. Otro problema que podemos encontrar es que los conjuntos de datos existentes no incluyen información sobre tipos de ataques emergentes, los cuales son de interés si se desea construir un sistema de gestión de ciberataques que sea capaz de manejar los tipos de ataques más actualizados y populares.

Para solucionar esta problemática se pretende construir un sistema que sea capaz de generar conjuntos de datos con ciberataques en redes virtuales. Este sistema se enfocará en simular, tanto tráfico benigno, como malicioso, empleando técnicas modernas de orquestación de contenedores como Docker y Docker Compose. La finalidad es que el sistema permita la creación de trazas de red (PCAP) y estadísticas etiquetadas de tráfico de red, mediante la simulación de ataques actuales que se han utilizado para generarlas. Estas trazas y estadísticas facilitarán una investigación actualizada en los sistemas de detección de ciberataques, basados en técnicas de inteligencia artificial.

De esta manera, para la generación de conjuntos de datos para la construcción de modelos de gestión de ciberataques, no será necesario hacer frente a un elevado presupuesto, ni será necesaria la implementación de una topología de red física, junto con una gran cantidad de equipos físicos para llevar a cabo simulaciones de ataques. De hecho, mediante la creación de una topología de red virtual dockerizada, sería posible replicar un entorno para realizar las simulaciones, realizando la captura y etiquetado de los datos. De esta manera, con el sistema propuesto, es posible obtener resultados en el simulador no muy diferentes a los obtenidos en un entorno compuesto por máquinas físicas, y todo ello, a un coste significativamente menor. Además, tendrá incluido un conjunto de los ataques más novedosos, e incluso será escalable permitiendo añadir, de manera muy simple, nuevos tipos de ataques.

Para ofrecer una mejor experiencia al usuario final, este sistema estará pensado para poder crear conjuntos de datos sin necesidad de tener conocimiento sobre cómo funcionan los diferentes ciberataques. Además, no se necesitará conocer las diferentes tecnologías a utilizar, puesto que el sistema estará pensado para realizarse de una manera cómoda y sencilla.

La estructura que tendrá el artículo es la siguiente: Sección II donde se describe el estado del arte; Sección III donde se describe cómo se ha construido el sistema generador de *datasets* con ciberataques, y Sección IV donde se definen los objetivos alcanzados y el trabajo futuro.

II. Estado del arte

La disponibilidad de conjuntos de datos públicos de alta calidad es fundamental para el avance en la investigación de Sistemas de Detección de Intrusiones (IDS). A pesar de la creciente importancia de este campo, el número de conjuntos de datos adecuados sigue siendo limitado. A continuación, se describen los conjuntos de datos más relevantes y utilizados en la comunidad científica:

- KDD Cup 1999[1]. Este conjunto de datos, derivado de DARPA98 y publicado por la Universidad de California, Irvine, en 1999, ha sido un pilar en la investigación de IDS. Su amplia adopción en el ámbito académico subraya su importancia histórica.
- NSL-KDD[2]. Desarrollado en 2009 como una mejora de KDD Cup 1999, NSL-KDD aborda varias deficiencias de su predecesor. Su objetivo principal fue corregir los problemas encontrados en el conjunto de datos anterior.
- UNSW-NB15[3]. Publicado por la Universidad de Nueva Gales del Sur (UNSW), este conjunto de datos ofrece una recopilación de tráfico de red que incluye una variedad de ataques modernos. Está estructurado en conjuntos de entrenamiento y prueba. Dispone de 9 tipos de ataques y 49 características.
- CIC-IDS-2017[4]. Reconociendo las limitaciones de los conjuntos de datos existentes, el Instituto Canadiense para la Ciberseguridad (CIC) desarrolló CIC-IDS-2017. Este conjunto de datos se destaca por su realismo y la inclusión de una amplia gama de ataques, lo que lo convierte en una referencia para la evaluación de modelos de detección de ciberataques.
- CSE-CIC-IDS2018[5]. Resultado de la colaboración entre CIC y la Organización de Seguridad de las Comunicaciones (CSE), CSE-CIC-IDS2018 se basa en la estructura de CIC-IDS-2017, con la distinción de haber sido generado en un entorno de Amazon Web Services (AWS). Ambos conjuntos de datos son considerados de los más modernos y actualizados.

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[2] https://www.unb.ca/cic/datasets/nsl.html
[3] https://research.unsw.edu.au/projects/unsw-nb15-dataset
[4] https://www.unb.ca/cic/datasets/ids-2017.html
[5] https://www.unb.ca/cic/datasets/ids-2018.html

- CIC-DDoS2019[6]. Publicado también por CIC, este conjunto de datos se centra específicamente en ataques de Denegación de Servicio Distribuido (DDoS), proporcionando una taxonomía detallada.
- AWID2[7] y AWID3[8]. Estos conjuntos, centrados en redes 802.11, ofrecen recursos valiosos para la investigación en seguridad de redes inalámbricas. AWID2 se centra en ataques comunes y sus firmas, mientras que AWID3 incluye ataques más modernos y complejos.
- UGR16 [9]. Publicado por la Universidad de Granada, este dataset contiene capturas realizadas durante 5 meses. Tiene 12 características y una amplia cantidad de datos.
- DAPT2020 [10]. Este dataset contiene una colección de capturas de 5 días. Dispone de las mismas características que los datasets publicados por CIC ya que procesaron capturas de ficheros PCAP, generadas por sus autores, con la herramienta CICFlowMeter.

La tabla I muestra esta revisión de los conjuntos de datos públicos disponibles, proporcionando una visión general de los recursos clave para la investigación en IDS. La selección del conjunto de datos adecuado dependerá de los objetivos específicos de cada estudio.

Dataset	Year	Attack Types	Features	Records
KDD Cup 1999	1999	4	41	4,898,431
NSL-KDD	2009	22	41	125,973
UNSW-NB15	2015	9	49	2,540,047
CIC-IDS-2017	2017	14	84	2,830,743
CSE-CIC-IDS2018	2018	14	84	16,232,943
CIC-DDoS2019	2019	18	84	12,794,627
AWID2	2016	3	155	42,388,298
AWID	2021	13	253	15,574,911
UGR16	2016	3	12	57,602,189
DAPT2020	2020	10	84	144,786

Tabla I
Conjuntos de datos existentes para la detección de ciberataques

El principal problema con los conjuntos de datos actuales es la escasez de conjuntos que cumplan unos requisitos mínimos de calidad como, por ejemplo, que contengan una gran cantidad de datos, los tipos de ataques más populares y modernos, y que se encuentren balanceados.

Para la generación de conjuntos de datos, una alternativa son las Redes Generativas Adverarias (GAN) ([11], [12], [13], [14]). Estos sistemas crean datos sintéticos de alta calidad lo más parecido a los datos originales. Una GAN está formada por dos redes neuronales, una que produce los datos sintéticos, y otra que trata de distinguir el dato sintético de los datos reales. De este modo asegura producir datos sintéticos con características similares a los de entrenamiento.

Los métodos basados en GAN para la generación de conjuntos de datos para IDS aunque resuelven el problema del desbalanceado presentan varias limitaciones entre ellas la imposibilidad de crear una topología de red sobre la que producir el dataset. Muchos estudios dependen de una combinación de datos reales y sintéticos para el entrenamiento, lo que puede introducir sesgos y limitar la generalización de los modelos IDS cuando se exponen a patrones de ataque nuevos o no vistos. Además, algunos estudios se centran en generar datos sintéticos basados en conjuntos de datos preexistentes, lo que puede restringir la variedad de escenarios de ataque simulados, ya que los datos generados pueden replicar únicamente patrones ya presentes en los conjuntos de datos originales, lo que puede llevar a que estos métodos no capturen amenazas IoT nuevas o en evolución. Muchos estudios no tienen en cuenta las demandas computacionales del procesamiento de datos en tiempo real, lo cual es crucial para los entornos IoT. Finalmente, la selección y optimización de características en métodos basados en GAN puede ser inconsistente, y la falta de selección puede llevar a ineficiencias.

Como alternativa, Vasilomanolakis et al. [15] propone la herramienta ID2T Toolkit, la cual solventa muchos de esos problemas. Las características son las siguientes:

- Crea conjuntos de datos etiquetados para evaluar IDS.
- Permite la generación de ataques definidos por el usuario, ofreciendo flexibilidad.
- Utiliza módulos para estadísticas, división de paquetes, control de ataques y fusión.
- Emplea generación basada en scripts y modificación de PCAP para generar ataques.
- Genera tráfico de fondo realista e inyecta diversos ataques de red y malware.

Sin embargo, los autores mencionan que se precisa intervención manual para las simulaciones. En este sentido, nuestro sistema, una vez configurado el entorno, no precisa de intervención alguna durante las simulaciones.

Por lo tanto, el sistema que proponemos solventa todas las carencias actuales que provocan que no se puedan generar nuevos conjuntos de datos con características deseables.

Otro de los aspectos los conjuntos de datos es la red sobre la que se realiza. Los autores de los conjuntos de datos CIC-IDS-2017 y CSE-CIC-IDS2018, generaron conjuntos de datos con una red real de equipos, y con una red en un entorno AWS, respectivamente. Ambas opciones requieren de recursos económicos elevados, mientras que con nuestro sistema, los costes siempre van a ser significativamente inferiores. Alternativamente se podría replicar el entorno mediante la creación de una red con Raspberries, que también requiere de recursos económicos para todo el equipamiento, pero dado el grado de madurez de los entornos virtuales, tomamos la decisión de crear esta herramienta para la generación de conjuntos de datos virtualmente, en nuestro caso con una infraestructura de máquinas con Docker. Esta herramienta permite que funcionen una gran cantidad de contenedores de manera conjunta, se puede construir una red o varias redes atacantes, e incluso una red o varias de víctimas, mediante el uso de Docker Compose. De este modo, todo se reduce a construir imágenes de máquinas, para su posterior instanciación como contenedores, los cuales cada uno de ellos tendrá un propósito distinto. Por ejemplo, un contenedor podría llevar a cabo ataques de fuerza bruta, otro ataques de inyección SQL, otro ataques de denegación de servicio, otro navega por internet, otro comprueba el correo, etc.

En los siguientes apartados explicaremos detalladamente cómo se ha construido todo este sistema para poder generar

[6]https://www.unb.ca/cic/datasets/ddos-2019.html
[7]https://icsdweb.aegean.gr/awid/awid2
[8]https://icsdweb.aegean.gr/awid/awid3

conjuntos de datos con ciberataques, además de las ventajas y desventajas que presenta.

III. Desarrollo del Sistema Generador de Conjuntos de Datos con Ciberataques

La propuesta que presentamos de un sistema para la generación de conjuntos de datos de ciberataques posee una serie de características deseables, como son las siguientes: (a) dispone de un conjunto de ciberataques ya incorporados en el sistema, a partir de una selección de entre los tipos de ataques más populares y actualizados hasta el momento; (b) hace uso de tecnologías relativamente novedosas cuya popularidad se encuentra en auge, como son los entornos virtualizados mediante contenedores virtuales, haciendo uso de la plataforma de virtualización Docker y Docker Compose, las cuales permiten implementar el diseño de una topología de red controlada, para la generación totalmente automatizada de simulaciones de ciberataques, de manera tan realista como si se intentara implementar en un entorno físico; (d) simplicidad ya que el usuario no necesitará conocer ninguna de las herramientas o tecnologías utilizadas para poder generar nuevos conjuntos de datos; y (e) se reducen significativamente los costes ya que solo es necesario un equipo informático modesto, en lugar del conjunto de equipos atacantes, víctimas, enrutadores, conmutadores, además del equipamiento eléctrico. En la Figura 1 se puede ser la arquitectura del sistema. Haciendo uso de la herramienta Docker, se crean contenedores dinámicamente con máquinas virtuales que ejecutarán las víctimas o atacantes. A partir de la información de red proporcionada, y haciendo uso de Docker Compose, se crea dinámicamente la topología de red especificada. Una vez creado todo el entorno se activarán los distintos equipos (víctimas y atacantes) durante el tiempo de duración de la simulación. Al finalizar producirá un fichero pcap con todas las trazas de la red, y etiquetando los paquetes que forman los ataques establecidos generando el fichero CSV con las estadísticas del flujo generado.

III-A. Ciberataques seleccionados para la ejecución de las simulaciones

La herramienta propuesta estará provista de una amplia selección de ciberataques, los cuales han sido minuciosamente seleccionados para que el sistema ya incluya una selección de los más populares y actualizados hasta el momento. Según distintas corporaciones enfocadas en la comercialización de soluciones enfocadas en la ciberseguridad, los ataques más comunes [16], [17], [18] de los últimos años son los siguientes: denegación de servicios (DoS), phishing, man-in-the-middle, inyección SQL, fuerza bruta, ataques hacia webs, ataques con algún tipo de malware, entre muchos otros.

Según el Centro de Estudios Estratégicos e Internacionales [19] (CSIS), los ataques que presentan mayor nivel de incidencia son: ataques de phishing o spear phishing, ataques con algún tipo de malware, ataques de denegación de servicio distribuido (DDoS), ataques de fuerza bruta, entre otros.

Para la selección de los tipos de ataques más populares y actualizados usaremos la matriz de MITRE ATT&CK como "base de conocimientos de acceso global sobre tácticas y técnicas de adversarios basadas en observaciones del mundo real"[9]. En esta matriz podemos observar diferentes tipos de técnicas de ataques cibernéticos, la definición de cada tipo, así como una lista de los distintos ataques que están recogidos por cada tipo de técnica junto con su definición. En los próximos puntos mencionaremos a qué tipos de técnicas corresponden nuestros ataques seleccionados. Uno de nuestros objetivos era incluir en el sistema al menos un ataque por cada tipo de técnica definida en la matriz de MITRE. Teniendo en cuenta que, para que la simulación de estos ataques sea de manera automática, se necesita de un conocimiento y herramientas muy específicos para replicarlos. Por lo que, se ha hecho una selección de distintas herramientas públicas para replicar los ciberataques de la mejor manera lo más realista posible.

De todas formas, es importante tener en cuenta que los tipos de ataques provistos en el sistema, son una selección de entre los más populares y actualizados. Esto no quiere decir que sea una selección cerrada ya que el sistema está preparado para que el usuario pueda añadir nuevos tipos de ataques, siempre y cuando tenga conocimientos suficientes para su replicación, además de conocimientos básicos de Docker.

Por lo tanto, una vez estudiados y seleccionados los tipos de ciberataques incluidos en el sistema propuesto, procederemos a definir cada uno de ellos. Estos ataques son: escaneo activo, fuerza bruta usando protocolo FTP, fuerza bruta usando protocolo SSH, fuerza bruta hacia una página web, DDoS hacia un servidor por protocolo HTTP, DDoS hacia una víctima por protocolo ICMP, DoS hacia un servidor por protocolo HTTP, DoS hacia una víctima por protocolo ICMP, man-in-the-middle, phishing, inyección SQL y troyano. Todos ellos hacen un total de 12 tipos de ciberataques diferentes.

III-A1. ***Escaneo activo:*** Este ataque, según la matriz de MITRE ATT&CK, está en las técnicas denominadas como "Reconocimiento"[10]. Un escaneo activo no es más que la búsqueda y análisis de puertos o puntos débiles en una red. Se procede a enviar paquetes con datos a diferentes puertos y se analizan las respuestas que se obtienen. Esto permite que los ciberdelincuentes encuentren posibles puntos de entrada que explotar, para futuros ataques que realicen[11].

Este ataque se ha seleccionado debido a que es un ataque muy común, que suele utilizarse junto a otro ataque. Dada su versatilidad y la frecuencia de su uso junto con los distintos ciberataques que existen, se ha escogido para nuestro sistema.

III-A2. ***Fuerza bruta:*** Se clasifica en las técnicas 'Acceso a credenciales"[12] según la matriz de MITRE ATT&CK. Para este ataque hemos utilizado tres tipos de variantes diferentes. Estas son fuerza bruta usando protocolo FTP, fuerza bruta usando protocolo SSH y fuerza bruta hacia una página web.

- **Fuerza bruta sobre protocolo FTP**: Este ataque consiste en probar sistemáticamente diferentes combinaciones de nombres de usuario y contraseñas, aprovechando que el sistema víctima atacado tenga alguna debilidad. Intenta adivinar las credenciales mediante "prueba y error", ya sea, de inicio de sesión, claves de cifrado, etc. Los ciberdelincuentes estudian todas las posibles combinaciones hasta conseguir acceso, en caso de que sea posible. Este ataque es efectivo si no se implementan las medidas de seguridad correctas como la limitación de

[9] https://attack.mitre.org/

[10] https://attack.mitre.org/techniques/T1595/

[11] https://www.fortinet.com/lat/resources/cyberglossary/what-is-port-scan

[12] https://attack.mitre.org/techniques/T1110/

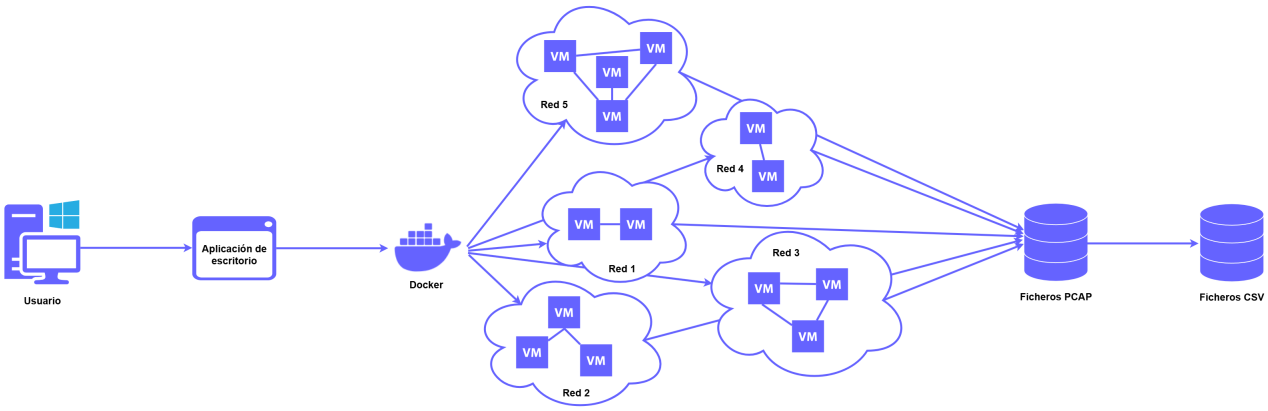


Figura 1. Gráfico de la arquitectura de nuestro sistema

intentos, la autenticación de doble factor, etc.

- **Fuerza bruta sobre protocolo SSH**: Este ataque es similar al anterior, pero utilizando SSH en lugar de FTP.
- **Fuerza bruta sobre Web**: En este tipo de ataque de fuerza bruta lo que se busca es intentar realizar una entrada en un formulario web. El atacante se encarga de enviar, como en cualquier ataque de fuerza bruta, repetidos intentos de inicio de sesión utilizando combinaciones de nombres de usuario y contraseñas intentando adivinar las credenciales de acceso. Al igual que en otros ataques de fuerza bruta mencionados anteriormente, la eficacia o no de este tipo de método depende del nivel de complejidad que tengan las contraseñas o las políticas de seguridad implementadas.

*III-A3. **Denegación de servicio distribuido**:* Los ataques de denegación de servicio distribuido (DDoS) se encargan de saturar la capacidad de respuesta de un servidor mediante el envío masivo de tráfico de red, en este caso utilizando solicitudes HTTP. Según la matriz de MITRE ATT&CK, este ataque se encuentra dentro del conjunto de técnicas llamado "Impacto"[13]. Para este ataque hemos utilizado dos tipos de variantes diferentes[14]: DDoS hacia un servidor por protocolo HTTP y denegación de servicio distribuido hacia una víctima por protocolo ICMP, de manera que se tienen en cuenta los típicos protocolos de navegación web y del correo electrónico.

- **DDoS sobre protocolo HTTP**: En este caso, los atacantes llevan a cabo una DDoS sobre HTTP enviando una gran cantidad de solicitudes al servidor de destino, haciendo que se ralentice considerablemente, o incluso se termine bloqueando. Este ataque concretamente se basa en un ataque de capa de aplicación. Cabe destacar que al ser un ataque distribuido se mandan paquetes desde diferentes máquinas.
  Este ataque se ha seleccionado debido a que su uso es muy común. Además, tiene un gran porcentaje de efectividad. Impidiendo el uso de sistemas, ya sean servidores u otras máquinas, por las víctimas que acceden a ellos.
- **DDoS sobre protocolo ICMP**: Al igual que el anterior tipo de ataque, este es un ataque DDoS pero que está basado en volumen, usando el protocolo ICMP. Esto quiere decir que este ataque inunda a la máquina víctima con paquetes de datos. Al sobrecargar la red de la víctima con tráfico ICMP, se consigue dificultar la comunicación de la víctima, causando que esta tenga interrupciones en los servicios y caídas en el sistema, provocando una no respuesta del mismo.
  Este ataque se ha seleccionado, al igual que el de capa de aplicación, debido a que su uso es muy común y ofrece una efectividad mayor que otros ataques. Y con esta variación de protocolo ICMP, en lugar de HTTP, podremos ver otro tipo de tráfico generado.

*III-A4. **Denegación de servicio**:* Según la matriz de MITRE ATT&CK, este ataque, como el ataque de DDoS, está dentro del conjunto de técnicas llamado "Impacto". A diferencia de un ataque de DDoS, un ataque de denegación de servicio (DoS) se origina desde una única máquina atacante. Para este ataque hemos utilizado dos tipos de variantes:

- **DoS sobre protocolo HTTP**: En este caso se usa un ataque basado en volumen. Este ataque es muy parecido al DDoS generando un envío constante y masivo de solicitudes HTTP a un servidor, agotando los recursos disponibles de este. Esto provoca lentitud o inclusive que se termine cayendo. Aunque este ataque suele ser más fácil de mitigar que un ataque distribuido, si no se toman las medidas necesarias, puede causar graves interrupciones en el servidor.
  Este ataque se ha seleccionado, al igual que los de DDoS, debido a que su uso es muy común y ofrece una efectividad mayor que otros ataques. Con la diferencia de que con este vamos a ver, en vez de un ataque distribuido, un ataque único. Además, utilizando el protocolo HTTP.
- **DoS sobre protocolo ICMP**: En este caso se usa un ataque de capa de aplicación. Este ataque es exactamente igual que el anterior ataque DoS, pero esta vez utilizando el protocolo ICMP para inundar a la víctima con una gran cantidad de paquetes de datos. Al sobrecargar la red de la víctima con tráfico ICMP, se consigue dificultar la comunicación de la víctima, causando que esta tenga

[13]https://attack.mitre.org/techniques/T1499/
[14]https://www.geeksforgeeks.org/types-of-dos-attacks/

interrupciones en los servicios y caídas en el sistema, provocando una no respuesta del mismo.
Este ataque se ha seleccionado, al igual que los de DDoS, debido a que su uso es muy común y ofrece una efectividad mayor que otros ataques. Con la diferencia de que con este vamos a ver, en vez de un ataque distribuido, un ataque único. Además, utilizando el protocolo ICMP.

III-A5. ***Man-in-the-middle****:* Según la matriz de MITRE ATT&CK, este ataque está dentro del conjunto de técnicas llamado "Recopilación"[15]. Este ataque también llamado en muchas ocasiones como "adversary-in-the-middle", consiste en interceptar y alterar la comunicación que existe entre 2 partes, sin que ninguna lo detecte. En un ataque MITM, las dos partes sienten que se están comunicando como lo hacen normalmente. Lo que no saben es que la persona que realmente envía el mensaje, modifica o accede ilícitamente al mensaje, antes de que llegue a su destino [16].

Este ataque se ha seleccionado debido a que es un ataque novedoso, como vimos anteriormente, que es utilizado en diversas situaciones en el día a día. Por lo que nos va a aportar bastante en el desarrollo de nuestro sistema.

III-A6. ***Phishing****:* Según la matriz de MITRE ATT&CK, pertenece a las técnicas denominadas"Acceso inicial"[16]. El phishing es una técnica en la que se envían correos electrónicos que parecen provenir de fuentes legítimas y confiables con el fin de obtener información confidencial de un objetivo. Los ataques de phishing combinan ingeniería social y tecnología y se denominan así porque el atacante, en efecto, está "pescando" el acceso a una zona prohibida utilizando el "cebo" de un remitente aparentemente confiable. Para ejecutar el ataque, el actor malicioso puede enviar un enlace que lo lleve a un sitio web que lo engañe para que descargue malware, como virus, o le proporcione al atacante su información privada. En muchos casos, es posible que la víctima no se dé cuenta de que ha sido atacada, lo que le permite al atacante atacar a otras personas de la misma organización sin que nadie sospeche de una actividad maliciosa [16].

Este ataque se ha seleccionado, al igual que el ataque anterior, debido a que es un ataque novedoso que es utilizado en diversas situaciones en el día a día. Además, podriamos decir que es de los más utilizados en la actualidad. Brindándonos de un gran aporte para nuestro sistema.

III-A7. ***Inyección SQL****:* Según la matriz de MITRE ATT&CK, este ataque está dentro del conjunto de técnicas llamado "Mando y control" [17]. La inyección SQL es un método múy común que explota las vulnerabilidades que tiene una aplicación web. El atacante intenta insertar, o inyectar, código SQL malicioso en campos de entrada. Lo que ocurre es que esta inyección que usa el atacante le permite sacar los datos de la base de datos del sistema, pudiendo acceder, modificar, o inclusive, eliminar información sensible. Ya que es prácticamente como si ejecutáramos una consulta SQL en la propia base de datos [16].

Este ataque se ha seleccionado porque, aunque sea un método de acceso de los ataques más antiguos, todavía sigue generando gran cantidad de brechas de seguridad, a causa de que las organizaciones pocas veces prestan atención en proteger sus sistemas de ataques de inyección SQL.

III-A8. ***Troyano****:* Según la matriz de MITRE ATT&CK, este ataque está dentro del conjunto de técnicas llamado "Ejecución"[18]. Un troyano, o caballo de troya, es un tipo de *malware* que se disfraza de un software, aparentemente legítimo, para engañar al usuario víctima logrando que este realice la instalación del programa. Una vez se ha instalado y ejecutado dicho programa malicioso, el troyano puede permitir al atacante desde acceder de forma remota al sistema infectado como robar información confidencial, instalar más programas maliciosos en el sistema de la víctima, o inclusive conseguir controlar el sistema. Todo esto produce que en muchas ocasiones, al ser algo que está ejecutándose normalmente en segundo plano, el usuario no es consciente de la presencia de este software malicioso en su equipo [16].

Este ataque se ha seleccionado dado su uso común y la flexibilidad que tiene. Además, es un tipo de ataque muy común en el que los ciberdelincuentes suelen tener éxito.

III-B. Implementación de los entornos virtuales

Una vez definidos en el apartado anterior, en qué consisten cada uno de los ciberataques seleccionados, procederemos a explicar cómo se ha conseguido simular estos ataques utilizando técnicas o herramientas públicas. Cabe destacar, que todos los ataques simulados se han encapsulado en imágenes de Docker. Estas imágenes, una vez instanciadas como contenedores, generarán el tráfico de red deseado, dependiendo del rol específico que tenga cada una de ellas, siendo los siguientes: atacante o víctima.

III-B1. Rol atacante: Encargado de ejecutar el ataque mediante scripts automatizados. Se emplean en él imágenes base (por ejemplo, Kali Linux o Ubuntu) junto con herramientas específicas que varían según el tipo de ataque, según se detalla a continuación:

- **Escaneo activo:** Se utiliza Nmap para identificar puertos abiertos en la máquina víctima, para posteriormente intentar acceder a dichos puertos.
- **Fuerza bruta (FTP, SSH, web):** Se emplea la herramienta Hydra y diccionarios para probar múltiples combinaciones de credenciales, ya sea para servicios FTP, conexiones SSH o formularios web.
- **Ataques de denegación de servicio (DoS/DDoS):** Se diferencian escenarios distribuidos y no distribuidos:
  - **DDoS hacia un servidor por protocolo HTTP:** Se utilizan múltiples contenedores atacantes que, coordinadamente, saturan un servidor web. Esto se conoce como un ataque de denegación de servicio de capa de aplicación.
  - **DDoS hacia una víctima por protocolo ICMP:** Emplea ataques de inundación ICMP (con Hping3) de forma distribuida para bloquer las acciones de la víctima. Esto se conoce como un ataque de denegación de servicio basado en volumen.
  - **DoS hacia un servidor o una víctima (HTTP e ICMP):** Se simulan ataques no distribuidos, en los que un único contenedor atacante ejecuta el ataque

[15]https://attack.mitre.org/techniques/T1557/
[16]https://attack.mitre.org/techniques/T1566/
[17]https://attack.mitre.org/techniques/T1659/
[18]https://attack.mitre.org/techniques/T1204/

hacía las diferentes víctimas (ya que si fuese más de una máquina atacante, sería un ataque distribuido).

- **Man-in-the-middle:** El atacante realiza diversas acciones para interceptar y modificar el tráfico entre una víctima y un servidor, utilizando herramientas como mitmproxy. además de scripts para gestionar las reglas de iptables que permiten la manipulación del tráfico de red entrante y saliente.
- **Phishing:** Se simulan ataques de phishing mediante herramientas como Zphisher y Expect. Además, se utilizan scripts en Python para el envío de correos fraudulentos.
- **Inyección SQL:** El atacante se encarga de realizar inyecciones de código SQL en formularios web mediante la herramienta sqlmap. Esto le permite al atacante extraer, modificar o eliminar datos de la base de datos de la víctima, aprovechando los fallos producidos en los campos de entrada del formulario.
- **Troyano:** Se implementa mediante la distribución, a la víctima, de software malicioso disfrazado de aplicaciones legítimas. El atacante utiliza scripts para automatizar la instalación y ejecución del troyano en el sistema de la víctima, permitiéndole acceder de manera remota y robarle información confidencial o controlar el sistema.

La automatización en cada uno de estos escenarios se gestiona mediante la configuración de cron y variables de entorno (como CRON TIMES, EXECUTION TIMES y VICTIM IPS), lo que permite definir de manera precisa la fecha de inicio, la duración y la frecuencia (repeticiones) de los ataques. Esto nos permite gestionar correctamente las máquinas que se creen.

III-B2. Rol víctima: Las máquinas víctimas se configuran, o bien, mediante Puppeteer, o a partir de imágenes extraídas de repositorios públicos (como Docker Hub o DockerLabs) y están seleccionadas para representar sistemas vulnerables. Además, se integran scripts que generan tráfico benigno (por ejemplo, navegación web simulada, lectura de correo electrónico, etc.) para complementarlo con la actividad maliciosa y generar trazas de red realistas.

III-B3. Rol servidor: En ataques distribuidos o en escenarios donde se simula un entorno cliente-servidor (como en algunos DDoS o en el ataque man-in-the-middle) se incluye un rol adicional. Este rol es el de servidor, el cual se basa en imágenes que proveen servicios web (por ejemplo, Nginx en el ataque de MITM) y se configura para mostrar contenido, el cual muchas veces, es revisado por las máquinas víctimas.

De los ataques definidos podemos obtener, para estos roles (víctimas y servidores), la siguiente información:

- **Escaneo activo:** La víctima se configura a partir de imágenes recogidas de Docker Hub. Estas máquinas simulan sistemas con servicios expuestos y sin medidas de seguridad avanzadas, permitiendo la detección de puertos abiertos. En este escenario, no se requiere una máquina específica con el rol servidor, ya que la víctima actúa como el único objetivo del escaneo.
- **Fuerza bruta (FTP, SSH, web):** La víctima dispone de servicios (FTP, SSH o aplicaciones web) configurados con credenciales que el atacante tratará de obtener mediante el uso de múltiples combinaciones de acceso. En los protocolos FTP y SSH se usa Puppeteer y se simulan navegaciones web benignas. En el caso del servicio web se utiliza una imagen de una máquina vulnerable, que no realiza ningún tipo de acción. En este ataque, tampoco se utiliza máquina servidor junto con las víctimas.
- **Ataques de denegación de servicio (DoS/DDoS):** En escenarios DDoS o DoS mediante protocolo HTTP, el servidor se implementa mediante imágenes que proveen servicios web (por ejemplo, Nginx). Donde las víctimas intentarán acceder al servidor, mediante navegación web, pero no podrán debido al ataque de DDoS o DoS producido por el atacante. Por otro lado, en escenarios mediante protocolo ICMP, se denegará completamente las acciones que intente realizar una víctima, ya sea navegación web, acceso al correo, etc. En este segundo caso no se utilizará el rol de servidor.
- **Man-in-the-middle:** La víctima se configura para establecer comunicaciones cotidianas con un servidor, sin mecanismos de seguridad avanzados que impidan la posible interceptación, de esa comunicación, por un tercero. Esto permite que el tráfico pueda ser capturado y modificado por el atacante. El servidor, por su parte, ofrece un contenido original, lo que facilita la posterior comparación entre la información legítima y la alterada durante el ataque.
- **Phishing:** La víctima simula a un usuario final que interactúa con correos electrónicos y sitios web, siendo vulnerable a acceder a un enlace fraudulento. Existe tambíen el rol de servidor, que será el que alamacene los correos enviados por la máquina atacante.
- **Inyección SQL:** La víctima se configura con la misma máquina web vulnerable que en el ataque de fuerza bruta web. Esta máquina es vulnerable a inyecciones de código SQL. Vulnerabilidad que aprovechará el atacante. En este ataque solo existe un serviro o víctima, y no los dos roles conjuntamente.
- **Troyano:** La víctima se prepara para simular un entorno de usuario final, donde se instalan aplicaciones que aparentan ser legítimas pero que contienen código malicioso. Luego de esto, la víctima realiza, mediante Puppeteer, navegaciones web benignas. En este ataque solo se utilza el rol de víctima.

Mediante la combinación de estas simulaciones, nuestro sistema permite la generación de conjuntos de datos con tráfico malicioso y benigno, que nos sirven para entrenar posteriormente a nuestros sistemas de detección basados en técnicas de inteligencia artificial, ofreciendo datos sumamente realistas, populares y actualizados que responden a las amenazas actuales que existen.

III-C. Construcción de la aplicación que generará los entornos virtuales

En esta sección se describen brevemente los puntos clave a destacar. Estos puntos son los siguientes:

- **Configuración de ataques predefinidos:** Esta opción permite seleccionar y configurar los ataques que deseen simular. Los usuarios pueden elegir entre una variedad de ataques predefinidos y ajustar las diferentes características que tienen cada uno de los diferentes ataques.
- **Visualización de la configuración actual:** Los usuarios, una vez han configurado o cargado algún ataque,

tienen la opción de poder ver la configuración actual de los ataques. Incluyendo detalles sobre las máquinas atacantes, víctimas y servidores, así como las fechas, duraciones y objetivos de las distintas máquinas.

- **Guardado y carga de las configuraciones:** La aplicación permite a los usuarios cargar configuraciones guardadas previamente desde ficheros JSON, así como guardar las configuraciones actuales, en ese mismo formato de ficheros, para su uso futuro.
- **Configuración detallada de cada ataque predefinido:** Los usuarios pueden configurar individualmente cada ataque, especificando el número de máquinas atacantes, víctimas y servidores (si estas existiesen), así como los detalles de cada máquina (nombre, dirección IP, máscara de subred, fechas, duraciones y repeticiones).
- **Generación de ficheros Docker Compose:** La aplicación permite generar automáticamente ficheros Docker Compose a partir de las configuraciones cargadas en la aplicación. Estos ficheros son necesarios para ejecutar los diferentes tipos de simulaciones.
- **Inicio y gestión de simulaciones:** Los usuarios pueden iniciar las simulaciones directamente desde la aplicación. Una vez iniciada alguna simulación, la aplicación proporciona información sobre el tiempo restante de la simulación y el procesamiento de los ficheros PCAP generados, para convertirlos en ficheros CSV.

IV. Conclusiones y trabajos futuros

En este artículo se ha presentado un sistema automatizado para la generación de conjuntos de datos con ciberataques, los cuales son los más populares y actualizados. Para la herramienta, se ha decidido usar entornos virtuales, mediante el uso de contenedores de Docker, junto con Docker Compose para la orquestación de toda la tipología de red diseñada, lo que permite simular tráfico benigno y malicioso, consiguiendo que el tráfico capturado contenga trazas de red sumamente realistas con una mezcla de tráfico de todo tipo. Los ataques seleccionados, están basados en amenazas actuales siguiendo la matriz MITRE ATT&CK, habiendo seleccionado los tipos de ataques más populares y actualizados hasta el momento. Además, se implementan mediante roles definidos (atacante, víctima, servidor). Todo ello se consigue mediante el uso de una herramienta que permite diseñar e implementar una topología de red, lo que permite que los conjuntos de datos generados sean sumamente realistas. Además, la herramienta implementada no requiere de conocimiento sobre cómo se implementan los ataques, ni cómo se utiliza Docker. Finalmente, se obtienen tanto la captura de tráfico de red, PCAP, además de ficheros CSV con estadísticas de tráfico de red.

Como trabajos futuros, debido al potencial de esta herramienta, se generarán y se publicarán nuevos conjuntos de datos, los cuales serán útiles para ayudar a los investigadores en una línea de investigación de vital importancia, como es la gestión de ciberataques, ya sea su detección o predicción. En cuanto al sistema, dada la amplia variedad de ciberataques existentes, se seleccionarán nuevos tipos de ataques para su inclusión dentro de la herramienta propuesta, lo que permitirá construir sistemas de gestión de ciberataques de calidad. Los tipos de ataques seleccionados tienen su propia naturaleza, además de sus propias peculiaridades, por lo que se ampliará la herramienta en cuanto al conjunto de opciones personalizables para incrementar, incluso más, el nivel de realismo de los ciberataques implementados. La herramienta permite modificar las imágenes para una mayor personalización, pero se ganaría en usabilidad si estas opciones se pudieran configurar gráficamente con la herramienta presentada.

Agradecimientos

Este trabajo ha sido posible gracias a la financiación del Plan Propio de la Universidad de Cádiz (Proyecto INERCIA PR2024-033), en el marco de la convocatoria de ayudas 2022-2023 para proyectos en áreas prioritarias.

Referencias

[1] W. S. Admass, Y. Y. Munaye, and A. A. Diro, "Cyber security: State of the art, challenges and future directions," 1 2024.

[2] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions," *SN Computer Science*, vol. 2, no. 3, p. 173, 2021.

[3] INCIBE, "¿qué son y para qué sirven los siem, ids e ips?" 2020. [Online]. Available: https://www.incibe.es/empresas/blog/son-y-sirven-los-siem-ids-e-ips

[4] M. Vielberth and G. Pernul, "A security information and event management pattern," in *12th Latin American Conference on Pattern Languages of Programs (SLPLoP)*, November 2018.

[5] A. S. Ashoor and S. Gore, "Difference between intrusion detection system (ids) and intrusion prevention system (ips)," in *Advances in Network Security and Applications: 4th International Conference, CNSA 2011, Chennai, India, July 15-17, 2011 4*. Springer, 2011, pp. 497–501.

[6] A. S. Ashoor and S. Gore, "Importance of intrusion detection system (ids)," *International Journal of Scientific and Engineering Research*, vol. 2, no. 1, pp. 1–4, 2011.

[7] Verizon, "2024 data breach investigations report," 2024. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/

[8] IBM, "Cost of a data breach report 2024," 2024. [Online]. Available: https://www.ibm.com/reports/data-breach

[9] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, and R. Therón, "Ugr'16: A new dataset for the evaluation of cyclostationarity-based network idss," *Computers & Security*, vol. 73, pp. 411–424, 2018.

[10] S. Myneni, A. Chowdhary, A. Sabur, S. Sengupta, G. Agrawal, D. Huang, and M. Kang, "Dapt 2020 - constructing a benchmark dataset for advanced persistent threats," in *Deployable Machine Learning for Security Defense*, G. Wang, A. Ciptadi, and A. Ahmadzadeh, Eds. Cham: Springer International Publishing, 2020, pp. 138–163.

[11] X. Liu, T. Li, R. Zhang, D. Wu, Y. Liu, and Z. Yang, "A gan and feature selection-based oversampling technique for intrusion detection," *Security and Communication Networks*, vol. 2021, no. 1, 2021.

[12] V. Kumar and D. Sinha, "Synthetic attack data generation model applying generative adversarial network for intrusion detection," *Computers & Security*, vol. 125, p. 103054, 2023.

[13] P. Ranka, A. Shah, N. Vora, A. Kulkarni, and N. Patil, "Computer vision-based cybersecurity threat detection system with gan-enhanced data augmentation," in *International Conference on Soft Computing and its Engineering Applications*. Springer, 2023, pp. 54–67.

[14] C. Strickland, M. Zakar, C. Saha, S. Soltani Nejad, N. Tasnim, D. J. Lizotte, and A. Haque, "Drl-gan: A hybrid approach for binary and multiclass network intrusion detection," *Sensors*, vol. 24, no. 9, 2024.

[15] E. Vasilomanolakis, C. G. Cordero, N. Milanov, and M. Mühlhäuser, "Towards the creation of synthetic, yet realistic, intrusion detection datasets," in *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2016, pp. 1209–1214.

[16] Fortinet, "Top 20 most common types of cyber attacks," 2025. [Online]. Available: https://www.fortinet.com/resources/cyberglossary/types-of-cyber-attacks

[17] Crowdstrike, "12 most common types of cyberattacks," 2024. [Online]. Available: https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/common-cyberattacks/

[18] Cisco, "What is a cyberattack?. most common types," 2025. [Online]. Available: https://www.cisco.com/c/en/us/products/security/common-cyberattacks.html

[19] CSIS, "Significant cyber incidents," 2025. [Online]. Available: https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents

Honeypot Adaptativa y Dinámica Impulsada por la IA para Estrategias de Engaño

Pedro Beltrán-López
Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia,
30100 Murcia, España
pedro.beltranl@um.es

Manuel Gil Pérez
Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia,
30100 Murcia, España
mgilperez@um.es

Pantaleone Nespoli
Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia,
30100 Murcia, España
pantaleone.nespoli@um.es

***Resumen*—En un contexto global marcado por la transformación digital y una interconexión sin precedentes, la superficie de ataque se ha expandido exponencialmente, lo que exige nuevas estrategias para defendernos de las amenazas emergentes y aprender de ellas. Integrando la Inteligencia Artificial (IA) con mecanismos de ciberengaño como los honeypots, es posible explotar el potencial combinado de ambos ámbitos, aún poco explorados. Este trabajo propone una innovadora arquitectura de honeypot potenciada por la IA, diseñada para adaptarse de manera dinámica a las tácticas de los atacantes. El sistema integra componentes fundamentales como la simulación del entorno, el procesamiento y la clasificación de los datos, la adaptabilidad en tiempo real y la orquestación de respuestas automáticas. Mediante técnicas de Procesamiento del Lenguaje Natural (NLP), en concreto la vectorización TF-IDF y el análisis de similitud del coseno, se ha desarrollado un mecanismo de recomendación que identifica el honeypot más adecuado a desplegar según la descripción del ataque. Los experimentos realizados demuestran un excelente rendimiento en diferentes escenarios.**

***Index Terms*—Ciberseguridad, Honeypot, Inteligencia Artificial, Procesamiento del Lenguaje Natural, Aprendizaje por Refuerzo, Adaptabilidad, Conciencia situacional**

Tipo de contribución: *Investigación original*

I. Introducción

En los últimos años hemos sido testigos de un aumento exponencial en el número de ataques y amenazas que afectan nuestros sistemas y entornos digitales [1]. Este incremento no solo se vincula con la expansión incesante de dispositivos conectados a Internet, sino también con la creciente sofisticación de las Técnicas, Tácticas y Procedimientos (TTPs) empleados por los atacantes para vulnerar y comprometer nuestros sistemas críticos [2]. Frente a esta escalada en la actividad maliciosa, resulta fundamental desarrollar e implementar estrategias innovadoras que permitan prevenir, detectar, reaccionar y analizar de manera proactiva cualquier intento de intrusión, asegurando la integridad de nuestras infraestructuras tecnológicas.

Una de las tendencias más prometedoras en el campo de la ciberseguridad es el ciberengaño [3], una técnica que no se limita a crear entornos falsos, sino que abarca un conjunto de estrategias diseñadas para confundir y desorientar a los atacantes. Su objetivo es atrapar a los atacantes en sistemas que parezcan reales proporcionándoles la información y las funcionalidades que estos desean sin exponer el sistema real, lo que permite tanto detectar ataques en sus primeras fases como analizar en profundidad sus métodos y tácticas [4]. Inicialmente basado en el uso de honeypots, i.e., sistemas creados para atraer y detectar a los atacantes [5], el concepto ha evolucionado hacia enfoques de engaño dinámico o defensa activa que se adaptan de manera continua y en tiempo real. Así, el ciberengaño sirve para desviar a los atacantes, dificultar la exfiltración de información y fortalecer las defensas mediante la retroalimentación obtenida del comportamiento del atacante. En consecuencia, ofrece la ventaja de dotar a las infraestructuras de mayor robustez y adaptabilidad frente a las amenazas emergentes, permitiendo a los defensores ajustar sus estrategias según el avance del ataque y la situación real de la red [6].

Paralelamente, el auge imparable de la Inteligencia Artificial (IA) y, en particular, de los modelos de lenguaje a gran escala (LLMs) [7], ha revolucionado numerosos sectores. Estas tecnologías han transformado la forma en que procesamos y analizamos grandes volúmenes de datos, aportando capacidades avanzadas de aprendizaje, razonamiento y generación de respuestas en tiempo real [8]. En el ámbito de la ciberseguridad, la integración de la IA resulta especialmente efectiva, ya que permite identificar patrones complejos frente a grandes cantidades de datos, anticiparse a comportamientos maliciosos y automatizar respuestas ante incidentes de seguridad con un nivel de precisión sin precedentes.

Para abordar este complejo desafío, la IA se posiciona como una herramienta indispensable, capaz de gestionar de manera eficiente la enorme cantidad de datos generados tanto por el entorno digital como por las actividades del atacante [9]. La sinergia entre ciberengaño y técnicas avanzadas de IA optimiza el análisis en tiempo real y permite una adaptación dinámica de las estrategias de defensa. Así, esta integración representa un hito fundamental en la evolución de la ciberseguridad, abriendo nuevas perspectivas para proteger de manera más efectiva las infraestructuras críticas ante amenazas cada vez más sofisticadas.

Diversos estudios recientes han asociado la IA a métodos tradicionales de ciberengaño, como el uso de honeypots, para mejorar la detección y el análisis de amenazas [10]. Sin embargo, la mayoría de estas investigaciones se centran en aspectos puntuales y dejan por explorar en profundidad su aplicación a lo largo de todas las fases del ataque. Además, es notable la escasez de trabajos que propongan una arquitectura integral capaz de incorporar de forma coherente y dinámica la

Ref	Año	IA	Adaptabilidad	Arquitectura	Finalidad	Escenario	Clasificación del atacante	Conciencia situacional	Experimentos
[11]	2017	ML	✓	✗	Desplegar	No definido	✗	✓	✓
[12]	2020	ML	✗	No definido	Detectar	IoT	✗	✗	✓
[13]	2023	LSTM	✗	✗	Detectar	No definido	✗	✗	✓
[14]	2023	LLM	✗	✗	Responder	No definido	✗	✗	✓
[15]	2023	No definido	✓	✓	Decidir Responder	No definido	✗	✗	✗
[16]	2024	LLM	✗	✗	Detectar	No definido	?	✗	✓
[17]	2024	LLM	✓	✗	Responder	Cloud	✗	✗	✓
[18]	2024	RL	✓	✗	Decidir	SDN	✗	✓	✓
Nuestro	2025	LLM GANs ML DL RL	✓	✓	Detectar Decidir Responder Simular Desplegar Aprender	Genérico	✓	✓	✓

Tabla I
COMPARACIÓN DE LOS TRABAJOS QUE RELACIONAN IA Y HONEYPOTS.

IA en los sistemas de honeypot, adaptándose continuamente al atacante. Esta limitación evidencia la necesidad de una aproximación integral que no solo incorpore la IA en los sistemas de honeypot, sino que además la utilice para adaptar y optimizar la respuesta en función del tipo de ataque.

En este artículo proponemos una arquitectura innovadora para honeypots impulsada por la IA, diseñada para ejecutar acciones de detección, adaptabilidad, respuesta, simulación y aprendizaje continuo en tiempo real.

Además, en respuesta a esta brecha, hemos desarrollado una prueba de concepto que valida la efectividad de integrar la IA para seleccionar dinámicamente el honeypot más adecuado, teniendo en cuenta el tipo de ataque y la conciencia situacional. Este enfoque no solo refuerza la adaptabilidad del sistema, sino que además permite una respuesta más precisa y oportuna ante amenazas emergentes.

A continuación, se presentan las principales contribuciones de este artículo:

- Creación y definición de una arquitectura integral para un honeypot potenciado por la IA, capaz de adaptarse dinámicamente a las acciones y estrategias de los atacantes con la finalidad de decidir, responder, simular y desplegar acciones en tiempo real para atrapar a un atacante el mayor tiempo posible.
- Identificación y descripción de los componentes fundamentales de dicha solución, junto con la implementación de mecanismos de IA que permitan la detección, clasificación y respuesta inmediata ante incidentes de seguridad.
- Desarrollo de una prueba de concepto con la que validar la efectividad de la integración de la IA para seleccionar dinámicamente el honeypot más adecuado en función del tipo de ataque y de la conciencia situacional.

El resto del artículo se organiza de la siguiente manera. En la Sección II se realiza una revisión y comparación detallada de los trabajos existentes en el estado del arte relacionados con el objetivo principal de este estudio. La Sección III presenta en profundidad la arquitectura del honeypot impulsada por IA, definiendo sus componentes y explicando los mecanismos de IA para garantizar su correcto funcionamiento y adaptabilidad. A continuación, la Sección IV describe la prueba de concepto llevada a cabo, que confirma la eficacia de la aplicación de técnicas de IA en entornos de ciberengaño. Finalmente, en la Sección V se exponen las conclusiones derivadas del estudio y se sugieren posibles líneas de investigación futuras para continuar avanzando en este campo tan crucial.

II. ESTADO DEL ARTE

La creciente sofisticación de las amenazas ha impulsado el desarrollo de enfoques innovadores que integran técnicas avanzadas para mejorar la detección, análisis y respuesta ante ataques, especialmente mediante el uso estratégico de honeypots. En este sentido, diversos estudios abordan el desafío desde múltiples ángulos y combinan herramientas de IA, computación en la nube y LLMs para lograr sistemas de engaño más realistas y efectivos. A continuación se hace un repaso de los principales estudios que abordan esta temática resumiendo sus principales características en la Tabla I. A continuación, definimos las principales características que hemos tenido en cuenta para comparar las soluciones encontradas en el estado del arte con nuestra propuesta:

- IA: Técnicas de IA utilizadas para modelar el comportamiento interno del honeypot.
- Adaptabilidad: El honeypot es capaz de modificarse automáticamente según las necesidades.
- Arquitectura: Proponen una arquitectura para el sistema de honeypot.
- Finalidad: Propósito del honeypot.
- Escenario: Escenario de aplicación del honeypot.
- Clasificación del atacante: Clasifican el tipo de ataque o atacante que está interactuando con el honeypot.
- Conciencia situacional: Es capaz de observar el entrono y obtener información para tomar decisiones.
- Experimento: Lleva a cabo experimentación en el artículo.

Como ejemplo, en [11] se propone un método automatizado basado en técnicas de Machine Learning (ML) que escanea la red para extraer características relevantes, tales como TCP-Stack, IP, MAC y servicios activos, y que utiliza una variante modificada del algoritmo k-means para agrupar sistemas con similitudes. Esto permite generar configuraciones específicas que imitan de forma convincente dispositivos reales, reduciendo la necesidad de intervención manual y aumentando la efectividad del engaño en entornos operativos.

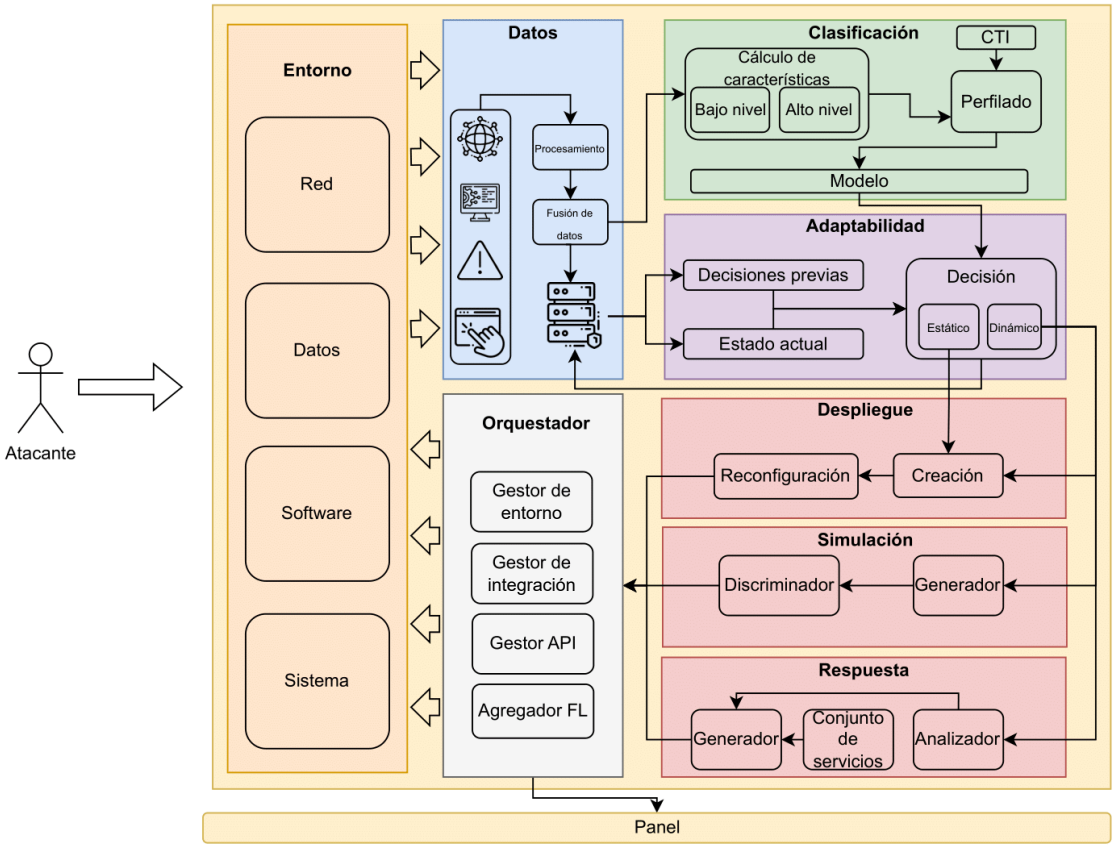


Figura 1. Diseño de honeypot adaptativa impulsada por la IA.

En otro enfoque, en [12], se integra IA con computación en la nube para detectar botnets en dispositivos del Internet de las Cosas (IoT). Utilizando técnicas de ML como la regresión logística, este sistema identifica ataques como ataques de denegación servicio distribuido (o DDoS, por sus siglas en inlgés) o ataques de intermediario (MitM), complementando las arquitecturas de seguridad existentes y dificultando la formación de redes de bots a gran escala.

Asimismo, en el trabajo [13], se ha desarrollado una solución enfocada en la detección de ataques SSH, basada en un honeypot que recoge registros heterogéneos (por ejemplo, mediante la implementación de Cowrie) y en un proceso de ingeniería de características que permite entrenar un modelo de memoria larga a corto plazo (LSTM) con una precisión cercana al 98 % en la clasificación de comportamientos maliciosos. Este método reduce significativamente la carga operativa sobre los analistas y agiliza la respuesta ante incidentes.

Por otro lado, la necesidad de profundizar en el estudio del comportamiento de los atacantes ha llevado a la creación de un honeypot interactivo que utiliza la API de ChatGPT en [14]. Esta solución simula un sistema vulnerable y, al responder a los comandos de los atacantes, recopila datos detallados sobre sus tácticas y estrategias, lo cual es fundamental para optimizar futuras soluciones de seguridad.

Además, el autor de [15] propone un marco basado en IA para el diseño de honeypots adaptativos y realistas. Este enfoque se estructura en tres componentes: un motor de análisis de comportamiento para comprender y predecir las acciones del atacante, un módulo de generación de entornos decoy que construye escenarios de engaño convincentes y un sistema de adaptación de respuestas que ajusta dinámicamente la interacción según la evolución del ataque. Los resultados obtenidos evidencian una mejora significativa en la capacidad para detectar y mitigar amenazas avanzadas, superando las limitaciones de los sistemas estáticos tradicionales.

Complementariamente, otro estudio se orienta a fortalecer la defensa mediante el uso de LLMs [16]. En este artículo los autores demuestran como el uso de LLMs es esencial para detectar y analizar amenazas ofreciendo una visión más compelta que con otras soluciones.

La incorporación de modelos de LLM también ha permitido desarrollar sistemas honeypot altamente interactivos en [17]. Mediante la recopilación y procesamiento de datos de registros, ingeniería de prompts y ajuste fino de modelos preentrenados, se logra simular el comportamiento de servidores Linux reales. Este sistema no solo genera respuestas contextuales y realistas durante las interacciones, sino que también mejora la capacidad de engaño y la calidad de la

información recopilada.

Finalmente, en [18] se ha implementado una estrategia que se centra en la detección de agentes de hacking basados en LLM. Esta estrategia combina técnicas activas, como la inyección de prompts para "secuestrar" metas o extraer instrucciones del sistema, con un análisis temporal de las respuestas. Dicho análisis permite distinguir entre agentes automatizados y otros actores, identificando así un número significativo de posibles ataques automatizados y aportando valiosa información sobre este nuevo tipo de amenaza.

El análisis de diversos estudios presentes en el estado del arte nos muestra la ausencia de una arquitectura o diseño genérico y compacto que utilice la IA no solo para la detección, sino también para tareas de adaptabilidad, simulación y respuesta. Este honeypot debe de tener en cuenta el contexto actual del entorno y del atacante para poder elegir de forma correcta las decisiones de despliegue, respuesta y simulación. Por ello, a continuación presentamos el diseño de un honeypot inteligente asistido con IA para llevar a cabo las tareas de engaño al atacante y obtener de manera inteligente y constante información de calidad.

III. Diseño de la solución

Después de identificar la ausencia de una arquitectura que integre el concepto de honeypot con IA, se propone la siguiente solución. En la Figura 1 se ilustra una arquitectura a alto nivel de un honeypot impulsada por IA, la cual es capaz de adaptarse de forma dinámica en tiempo real ante las acciones del atacante. A continuación, se describen y explican en detalle cada uno de sus componentes.

Para que el honeypot cuente con la capacidad de adaptarse de manera dinámica, es fundamental integrar diversos componentes que permitan recopilar, procesar y analizar una gran cantidad de información en tiempo real. Estos elementos trabajan conjuntamente para determinar la acción más prometedora en cada instante, de modo que el atacante no perciba que está siendo engañado y, al mismo tiempo, se obtenga información valiosa sobre sus acciones. Los componentes principales de esta arquitectura son los siguientes:

- **Entorno:** Este es uno de los componentes más críticos, ya que su función principal es mantener al atacante dentro del honeypot. El entorno simula un sistema real mediante la integración de múltiples fuentes de datos y archivos que emulan información acorde al tipo de ataque y al perfil del atacante (por ejemplo, trazas de llamadas a sistema, capturas pcap, listas de procesos, claves de registro y artefactos comunes en objetivos reales). Además, establece comunicación a nivel de red con otros entornos simulados para evitar la sensación de aislamiento. Para generar ruido de fondo creíble (como modificaciones al sistema de archivos o tráfico saliente legítimo) se entrenan modelos generativos (por ejemplo, modelos basados en transformadores o VAE) sobre grandes colecciones de huellas de host y de red recopiladas de sistemas de producción desinfectados. El uso de IA es clave porque permite sintetizar actividad coherente y contextual que evoluciona en tiempo real, logrando que el atacante perciba un sistema vivo y no un señuelo estático. Se configura con un sistema operativo y diversos servicios o aplicaciones que respondan a las necesidades detectadas en el atacante, adaptándose mediante plantillas parametrizables que la IA puede instanciar dinámicamente. La misión principal es engañar al atacante y lograr que permanezca interactuando en el entorno simulado con datos, redes, software y sistemas.
- **Datos:** En cualquier sistema basado en IA, el manejo de datos es esencial. Este componente se encarga de recopilar la información generada en las diversas dimensiones de actuación del entorno (tráfico de red, comandos introducidos, cambios en el sistema de archivos, artefactos de memoria, syscalls, logs de aplicaciones, métricas de latencia y contexto temporal) mediante agentes especializados desplegados a nivel de host y/o red. Los datos brutos se almacenan en un data lake; posteriormente pasan por un proceso de limpieza (normalización de campos, anonimización de direcciones IP internas, eliminación de duplicados) y preprocesamiento, donde se extraen features relevantes (n-gramas de comandos, histogramas de puertos, secuencias de llamadas). Se aplica fusión de información para crear una línea de tiempo única y granular que permite la trazabilidad completa de los eventos. Mantener este histórico estructurado posibilita entrenar modelos supervisados (clasificación) y no supervisados (detección de anomalías) con datos bien etiquetados, asegurando decisiones más efectivas y adaptaciones consistentes en fases posteriores.
- **Clasificación:** Antes de la fase de clasificación, se realiza un cálculo exhaustivo de características a partir de los datos capturados por el honeypot, combinando señales procedentes de fuentes de información internas (logs, trazas de red, configuraciones) y externas (feeds de CTI, bases de datos públicas sobre amenazas). Una vez que estos datos han sido preprocesados y fusionados, el componente de clasificación extrae las características esenciales que permiten identificar tanto el tipo de atacante como el tipo de ataque en curso. Para ello se emplea un pipeline de aprendizaje supervisado que combina algoritmos de embedding (Word2Vec para comandos, autoencoders para trazas de red) con modelos de decisión de alto rendimiento (Gradient Boosting, Random Forest o arquitecturas de redes profundas) entrenados con datasets etiquetados según la matriz MITRE ATT&CK y perfiles APT. El uso de IA está plenamente justificado: la diversidad y el gran volumen de señales hacen inviable la clasificación manual. Además, las técnicas de aprendizaje permiten reconocer patrones sutiles que cambian con el tiempo. El resultado alimenta la siguiente capa indicando la severidad, el objetivo probable y la motivación del adversario.
- **Adaptabilidad:** Con la identificación del atacante y del ataque, este componente decide las acciones a tomar. La primera vez que se detecta un flujo malicioso a nivel de red, un modelo contextual multi-brazo (bandits) selecciona la variante de honeypot que maximiza la probabilidad de interacción, basándose en variables como vector de entrada, origen geográfico y reputación de la IP (estático). A partir de la segunda interacción (dinámico), un agente de aprendizaje por refuerzo profundo (Deep RL),

entrenado sobre secuencias históricas de <estado, acción, recompensa>, ajusta dinámicamente los servicios, banners, puertos expuestos y los contenidos señuelo. Se emplean datos como duración de sesión, comandos por minuto y profundidad de directorios accedidos para definir la función de recompensa. Cuanto más tiempo y más comandos significativos ejecute el atacante, mayor recompensa.

- **Despliegue:** Este componente ejecuta las acciones de reconfiguración, creación de servicios y cualquier otro cambio decidido por la capa de Adaptabilidad. Se apoya en infrastructure-as-code (Terraform, Ansible) para reproducir de forma determinista imágenes de contenedores, reglas de cortafuegos y topologías virtuales.
- **Simulación:** La simulación es responsable de recrear interacciones y tráfico realista dentro del honeypot. Utilizando redes generativas adversariales (GANs) entrenadas con capturas de tráfico legítimo y de ataque de grandes honeynets universitarias e industriales, el componente genera flujos sintéticos (DNS, HTTP/2, SMB, llamadas RPC) que respetan distribuciones de tamaño de paquete, tiempo entre paquetes y patrones de retransmisión. Además, se entrena un modelo secuencial (Transformer) con playbooks de usuarios reales para simular comandos rutinarios en segundo plano. La IA permite así producir variabilidad infinita con coherencia estadística, dificultando que el atacante detecte repeticiones o anomalías que delatarían el entorno [19].
- **Respuesta:** Basado en mecanismos automatizados, este componente genera respuestas coherentes y creíbles frente a las acciones del atacante en situaciones críticas. Se emplean LLMs entrenados con conjuntos de datos de interacciones de servicios comunes (salidas de ssh, ps, netstat, mysql, APIs REST), reforzados con técnicas para optimizar la verosimilitud. Los datos de entrenamiento incluyen transcripciones de sesiones en laboratorios de ciberseguridad y volcados de terminales reales (anonimizados). La IA es crucial porque ofrece variaciones infinitas de respuestas manteniendo coherencia contextual, evitando plantillas estáticas que un atacante experto podría reconocer.
- **Orquestador:** Finalmente, el orquestador integra y coordina todos los componentes del sistema. Centraliza la toma de decisiones de alto nivel, alimentado por un modelo de gráficas de conocimiento que relaciona actores, técnicas y eventos del honeypot, construido a partir de los metadatos agregados anteriormente. Un submódulo de federated learning distribuye y agrega pesos de los modelos de clasificación y adaptabilidad entre instancias de honeypots, usando datos agregados (gradientes) y preservando la privacidad de los registros crudos. El dashboard consume estos metadatos y ofrece insights explicables generados por un modelo de razonamiento neuro-simbólico, justificando por qué se eligió una variante concreta o qué indicadores llevaron a una adaptación. El uso de IA aquí garantiza coherencia global, evita silos de conocimiento y acelera la evolución del sistema ante nuevas TTPs sin comprometer la confidencialidad de cada despliegue.

IV. Experimentación

Una vez diseñada la arquitectura de nuestro honeypot impulsado por la IA, se llevó a cabo un experimento para evaluar la efectividad de la intersección entre la IA y los honeypots en la identificación y asignación óptima de estos sistemas de defensa según el tipo de ataque detectado. En particular, se buscó determinar qué honeypot es la más adecuada para engañar al atacante en función de las características de su actividad maliciosa.

IV-A. Construcción del Conjunto de Datos de Honeypots

Para desarrollar un modelo capaz de realizar recomendaciones efectivas, primero recopilamos un conjunto de datos que contiene información relevante sobre los honeypots más utilizados en la actualidad [22]. Este proceso de recolección de datos incluyó la extracción de información desde diversas fuentes especializadas en ciberseguridad, documentación técnica y bases de datos abiertas sobre honeypots. El conjunto de datos final incluye las siguientes características clave de cada honeypot:

- Descripción breve: Resumen general del honeypot.
- Protocolos soportados: Tecnologías y estándares que el honeypot puede emular.
- Ataques mitigados: Tipos de amenazas que puede detectar o neutralizar.
- Notas adicionales: Información relevante sobre su implementación o funcionamiento.

Dado que la cantidad de datos estructurados disponibles sobre honeypots es limitada, cada entrada de nuestro conjunto de datos representa una instancia única con su información contextual específica. Esta baja disponibilidad de ejemplos etiquetados imposibilita la aplicación de enfoques tradicionales de aprendizaje supervisado, que requieren grandes volúmenes de datos para entrenar modelos con capacidad de generalización. Por este motivo, descartamos el uso de sistemas expertos, cuya dependencia de reglas fijas dificulta su adaptación a nuevas amenazas y exige un mantenimiento constante por parte de expertos humanos. En su lugar, optamos por técnicas NLP [20], que nos permiten evaluar la similitud semántica entre las descripciones textuales de los ataques y las características operativas de los honeypots. Esta estrategia no solo supera las limitaciones impuestas por la falta de anotaciones, sino que también ofrece una solución más flexible y escalable para entornos en constante evolución.

IV-B. Vectorización y Representación del Conocimiento

Para permitir el análisis y comparación de texto, transformamos las descripciones de honeypots en representaciones numéricas utilizando el modelo frecuencia de término – frecuencia inversa de documento (TF-IDF) [24]. Este método nos permite asignar pesos a cada término en función de su importancia relativa dentro del corpus de texto, reduciendo la influencia de palabras comunes y destacando términos distintivos. El proceso de vectorización consistió en los siguientes pasos:

1. Concatenación de características: Se combinaron las columnas de “Descripción Breve”, “Protocolos soportados”, “Ataques Mitigados” y “Notas Adicionales” en una única representación textual para cada honeypot.

Fichero	Descripción
attack_descriptions_clear	Detectamos múltiples intentos de fuerza bruta en conexiones SSH con descarga de malware
attack_descriptions_ambiguous	Se observan variaciones en el tráfico DNS, con picos y consultas inusuales que pueden estar relaionados con botnets
attack_descriptions_poor	Se detecta actividad anómala en el tráfico de red, con picos en conexiones y algunos errores leves

Tabla II
EJEMPLOS DE ATAQUES DESCRITOS EN LOS FICHEROS DE PRUEBA

2. Aplicación de TF-IDF: Se convirtió el conjunto de descripciones en una matriz de vectores numéricos de alta dimensionalidad.
3. Cálculo de similitudes: Se empleó la métrica de similaridad del coseno para medir el grado de correspondencia entre las descripciones vectorizadas [25].

IV-C. Mecanismo de Recomendación de Honeypots

Desarrollamos un sistema de recomendación que, dada una descripción de un ataque, retorna los honeypots más relevantes ordenados por similitud semántica. El proceso de recomendación sigue estos pasos:

1. Se recibe una descripción del ataque. En la Tabla II se observa un ejemplo de descripción de los ataques en cada uno de los ficheros. Los ataques que se han utilizado son SQL Injection, DNS Tunneling, Code Injection, Malware Delivery, Web exploit, Brute Force, DoS, entre otros.
2. Se transforma la descripción en un vector TF-IDF utilizando el mismo modelo entrenado con el conjunto de datos de honeypots.
3. Se calcula la similitud del coseno entre la descripción del ataque y cada una de los honeypots almacenados.
4. Se ordenan los honeypots según el índice de similitud, seleccionando las N mejores coincidencias.

Este enfoque permite la identificación de los honeypots que son semánticamente más adecuados para interactuar con atacantes según el tipo de ataque detectado.

IV-D. Evaluación del Sistema de Recomendación

Para evaluar el desempeño del sistema, se diseñó un conjunto de pruebas utilizando descripciones de ataques con distintos niveles de claridad. Este conjunto de pruebas se compuso a partir de tres archivos con características diferenciadas (1090 filas): uno que contenía descripciones detalladas y bien definidas (*attack_descriptions_clear.csv*), otro que incluía descripciones vagas o incompletas (*attack_descriptions_ambiguous.csv*) y un tercero que presentaba descripciones con información mínima o términos confusos (*attack_descriptions_poor.csv*). Cada uno de estos archivos incluía, por cada entrada, tanto la descripción del ataque como el honeypot esperado, es decir, aquel que idealmente debería ser recomendado por el sistema. Los archivos se generaron mediante un proceso sistemático que implicó la extracción de datos de fuentes especializadas en ciberseguridad y documentación técnica, seguido de una clasificación basada en las principales características de los honeypots utilizados. La evaluación se basó en métricas estándar, donde se verificó la exactitud en niveles top-N (analizando si el honeypot esperado se encontraba dentro de las N mejores recomendaciones) y se generó un reporte de clasificación enfocado en la predicción top-1. Es decir, nosotros tenemos un dataset con el honeypot esperado para cada descripción de ataque y nuestro modelo predice 10 honeypots a usar en orden. Si ese honeypot está dentro de esas 10, forma parte del top-10, si está dentro de las 5 primeras forma parte del top-5 y así sucesivamente.

El proceso de evaluación consistió en aplicar el sistema a cada descripción de ataque, comparar la recomendación obtenida con el honeypot esperado y calcular la exactitud en cada uno de los niveles evaluados.

IV-E. Análisis de Resultados del Experimento

Tras ejecutar el experimento con el conjunto de datos de honeypots y los archivos que contienen las descripciones de ataques (por ejemplo, ataques de fuerza bruta, DoS, spoofing, entre otros), obtuvimos los resultados que se muestran en la Tabla III. Esta tabla muestra la cantidad de predicciones acertadas según el archivo y el ranking donde se encontraba el honeypot correcta. A continuación, analizamos el desempeño del sistema de recomendación en función de la claridad de las descripciones de ataque.

Para visualizar de forma intuitiva cómo se relacionan los términos más significativos con cada honeypot, se generó el mapa de calor que se presenta en la Figura 2. En este gráfico, el eje horizontal representa los términos extraídos del corpus, mientras que el eje vertical agrupa los diferentes honeypots. La intensidad del color (con tonalidades más intensas en rojo) indica el peso o frecuencia relativa de cada término, permitiendo identificar de forma inmediata las áreas de mayor correspondencia. Con esta visualización se observa, por ejemplo, que algunos honeypots destacan en columnas asociadas a ataques HTTP, lo cual respalda la asignación realizada por el sistema de recomendación. A continuación, se detallan los hallazgos en función de la claridad de las descripciones de ataque:

- Descripciones Claras: Cuando se usaron descripciones detalladas y bien definidas, el sistema alcanzó un desempeño muy alto. Esto se refleja en que la exactitud fue del 96 % en el Top-1, mientras que el Top-3, Top-5 y Top-10 alcanzaron un 100 %.
- Descripciones Ambiguas: Con descripciones vagas (aquellas que resultan imprecisas y utilizan términos generales que pueden llevar a interpretaciones múltiples, como “problema en el sistema” sin detalles específicos) o incompletas (que, aunque precisas en lo mencionado, omiten información esencial necesaria para comprender la totalidad del contexto), se observó una disminución en el desempeño en la recomendación de la mejor opción: la exactitud en Top-1 cayó al 70,4 %, aunque las métricas Top-3, Top-5 y Top-10 subieron desde el 75,5 % hasta el 84,7 %. Esto indica que, aunque la recomendación en la posición más alta falla en la mitad de los casos, la opción correcta aparece en las listas extendidas.



Figura 2. Mapa de calor de la relación de las palabras más importantes con cada uno de los honeypots en los documentos de prueba.

Fichero	Top-1	Top-3	Top-5	Top-10
attack_descriptions_clear	1056/1090	1090/1090	1090/1090	1090/1090
attack_descriptions_ambiguous	768/1090	824/1090	884/1090	924/1090
attack_descriptions_poor	259/1090	348/1090	559/1090	790/1090

Tabla III
RESULTADOS DEL SISTEMA DE RECOMENDACIÓN CON DIVERSOS CONJUNTOS DE PRUEBA.

- Descripciones Deficientes: En el caso extremo de descripciones mínimas o mal estructuradas, el sistema mostró una reducción generalizada del rendimiento, con una exactitud del 23,7 % en el Top-1, mientras que los demás niveles se mantuvieron entre el 31,9 % y el 72,4 %.

Los resultados obtenidos en este experimento destacan la alta efectividad del sistema en condiciones ideales y ponen en valor la eficiencia de la IA para asignar honeypots. Los hallazgos subrayan que la integración de técnicas de NLP y modelos de recomendación basados en similitud semántica ofrece un gran potencial para optimizar la defensa en ciberseguridad, adaptándose de forma precisa a las características de los ataques. Además, este estudio constituye una prometedora prueba de concepto que evidencia la capacidad de la IA para potenciar las estrategias de honeypots.

V. CONCLUSIONES

El ciberengaño asistido por IA se posiciona como una estrategia vanguardista que, además de fortalecer las defensas de los sistemas, permite extraer inteligencia valiosa mediante el engaño activo de los atacantes en tiempo real. Este trabajo presenta una arquitectura integral para honeypots novedosa impulsada por la IA en la que somos capaces de integrar diversos mecanismos de IA para clasificar y ajustar dinámicamente el honeypot con el que el atacante está interactuando. Además, se ha realizado un sofisticado sistema de recomendación basado en técnicas de NLP y vectorización TF-IDF. Dicho sistema es capaz de seleccionar dinámicamente el honeypot más adecuado en función del tipo de ataque y la conciencia situacional, lo que garantiza una respuesta oportuna y eficaz ante amenazas emergentes.

El diseño de la solución destaca por su orquestación inteligente, que coordina de forma precisa la recopilación, procesamiento y análisis de datos en tiempo real. Cada componente (desde la simulación del entorno hasta la clasificación del atacante y la adaptabilidad del sistema) trabaja de manera integrada para asegurar que el atacante permanezca inmerso en un entorno creíble y dinámico. La implementación del componente de respuesta automatizada, apoyado en modelos de lenguaje, permite generar interacciones realistas que refuerzan la credibilidad del engaño, mientras que el uso de modelos de aprendizaje por refuerzo optimiza continuamente las decisiones del sistema.

En cuanto a los resultados experimentales se destaca que, en condiciones óptimas con descripciones de ataques claras, el sistema alcanzó una exactitud del 96 % en el nivel Top-1 y un 100 % en los niveles Top-3, Top-5 y Top-10, con métricas de precisión, recall y F1-score de 1,00. Estos resultados numéricos resaltan la efectividad de la solución cuando se dispone de información completa y precisa, evidenciando el gran potencial del enfoque propuesto para la optimización de sistemas de defensa en ciberseguridad. Como trabajos futuros están la implementación de los distintos bloques propuestos en

la arquitectura y la generación de modelos de IA que puedan modificar en tiempo real el comportamiento del honeypot y atrapar al atacente.

Agradecimientos

Este trabajo ha sido financiado parcialmente por el proyecto estratégico DEFENDER del Instituto Nacional de Ciberseguridad (INCIBE) y por el plan de Recuperación, Transformación y Resiliencia de los fondos Next Generation EU.



Este trabajo también ha recibido financiación del Fondo Europeo de Defensa (EDF) en virtud del acuerdo de subvención nº 101103044 EDF-2021-CYBER-R-CDAI-2.

Descargo de responsabilidad: No obstante, los puntos de vista y opiniones expresados son exclusivamente los de los autores y no reflejan necesariamente los de la Unión Europea (UE). La UE no se hace responsable de ellas.

Referencias

[1] Alshamrani, A., Myneni, S., Chowdhary, A., & Huang, D. (2019). "A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities". *IEEE Communications Surveys & Tutorials*, 21(2):1851–1877.

[2] The MITRE Corporation (2025). "MITRE ATT&CK". Disponible en: https://attack.mitre.org

[3] Beltrán-López, P., Gil Pérez, M., & Nespoli, P. (2024). "Cyber deception: State of the art, trends and open challenges". *arXiv preprint* arXiv:2409.07194.

[4] Beltrán-López, P., Nespoli, P., & Gil Pérez, M. (2024). "Cyber deception powered by artificial intelligence: Overview, gaps, and opportunities". In *IX Jornadas Nacionales de Investigación En Ciberseguridad* (pp. 254-261).

[5] Qin, X., Jiang, F., Cen, M., & Doss, R. (2023). "Hybrid cyber defense strategies using honey-X: A survey". *Computer Networks*, 230:109776.

[6] Guo, F., Feng, H., Qiu, S., Yu, Z., Li, J., & Tong, Y. (2024). "Hierarchical network security situational awareness model based on honeypot technology". In: *Intelligent Computing Technology and Automation*, pp. 199–206.

[7] Zhang, J., Bu, H., Wen, H., Liu, Y., Fei, H., Xi, R., & Meng, D. (2025). "When LLMS meet cybersecurity: A systematic literature review". *Cybersecurity*, 8:55.

[8] Beltrán-López, P., Nespoli, P., & Gil Pérez, M. (2024). Cyber deception reactive: TCP stealth redirection to on-demand honeypots. arXiv preprint arXiv:2402.09191.

[9] Tahseen, A., Shailaja, S.R., & Ashwini, Y. (2024). "Extraction for big data cyber security analytics". *Advances in Computational Intelligence and Informatics: Proceedings of ICACII 2023*, 993:365.

[10] Sun, C., Bu, Y., Chen, B., Zhang, D., Chen, Z., Lu, X., & Sun, J. (2022). "Application of artificial intelligence technology in honeypot technology". In: *2021 International Conference on Advanced Computing and Endogenous Security*, pp. 1–9.

[11] Fraunholz, D., Zimmermann, M., & Schotten, H.D. (2017). "An adaptive honeypot configuration, deployment and maintenance strategy". In: *2017 19th International Conference on Advanced Communication Technology*, pp. 53–57.

[12] Memos, V.A., & Psannis, K.E. (2020). "AI-powered honeypots for enhanced IoT botnet detection". In: *2020 3rd World Symposium on Communication Engineering*, pp. 64–68.

[13] Arshad, T., & Menon, S. (2023). "AI-enabled honeypot". *Journal of Network & Information Security*, 11(2):17–29.

[14] Raut, U., Nagarkar, A., Talnikar, C., Mokashi, M., & Sharma, R. (2023). "Engaging attackers with a highly interactive honeypot system using ChatGPT". In: *2023 7th International Conference On Computing, Communication, Control And Automation*, pp. 1–5.

[15] Ivanov, Y. (2023). "AI in cyber deception: Creating adaptive and realistic honeypot systems". *Advances in Computer Sciences*, 6(1):1–9.

[16] Volkov, D. (2024). "LLM agent honeypot: Monitoring AI hacking agents in the wild". *arXiv preprint arXiv:2410.13919*.

[17] Otal, H.T., & Canbaz, M.A. (2024). "LLM honeypot: Leveraging large language models as advanced interactive honeypot systems". In: *2024 IEEE Conference on Communications and Network Security*, pp. 1–6.

[18] Gao, D., Guo, C., Zhang, Y., Ji, W., Lv, Z., Li, Z., & Lin, R. (2024). "Risk-aware SDN defense framework against anti-honeypot attacks using safe reinforcement learning". *International Journal of Network Management*, 34(6):e2297.

[19] Kholgh, D. K., & Kostakos, P. (2023). PAC-GPT: A novel approach to generating synthetic network traffic with GPT-3. *IEEE Access*, 11, 114936-114951.

[20] Ismail, W.S. (2024). "Threat detection and response using AI and NLP in cybersecurity". *Journal of Internet Services and Information Security*, 14(1):195–205.

[21] Yinka-Banjo, C., & Ugot, O.A. (2020). "A review of generative adversarial networks and its application in cybersecurity". *Artificial Intelligence Review*, 53:1721–1736.

[22] paralax (s.f.). "Awesome Honeypots". Disponible en: https://github.com/paralax/awesome-honeypots

[23] Islam, M.Z., Ali, R., Haider, A., Islam, M.Z., & Kim, H.S. (2021). "Pakes: A reinforcement learning-based personalized adaptability knowledge extraction strategy for adaptive learning systems". *IEEE Access*, 9:155123–155137.

[24] Bafna, P., Pramod, D., & Vaidya, A. (2016). "Document clustering: TF-IDF approach". In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques*, pp. 61–66.

[25] Xia, P., Zhang, L., & Li, F. (2015). "Learning similarity with cosine similarity ensemble". *Information Sciences*, 307:39–52.

Sesión Investigación VII: Plataformas y marcos de defensa cibernética

Hacia una ciberseguridad holística en e-Gobierno: Extensión de X-Road y MARISMA en el framework GAUCHO

Víctor Figueroa
GSyA Research Group.
Universidad de Castilla-La Mancha,
Neuquén, Argentina
romualdovictor.figueroa@alu.uclm.es

Luis E. Sánchez Crespo, Antonio Santos-Olmo,
David G. Rosado, Eduardo Fernández-Medina
GSyA Research Group,
Universidad de Castilla-La Mancha, Ciudad Real, España
{luise.sanchez, antonio.santosolmo, david.grosado, eduardo.fdezmedina}@uclm.es

***Resumen*—La creciente digitalización de los servicios gubernamentales exige abordar la ciberseguridad de forma holística, protegiendo la información sensible y facilitando la interoperabilidad segura entre los sistemas que sustentan dichos servicios. Este artículo presenta un enfoque integral para e-Gobierno, basado en el framework *GAUCHO*, que integra tres componentes: la interoperabilidad segura de datos con X-Road, la gestión adaptativa de riesgos con MARISMA y la Gobernanza Estratégica de la Ciberseguridad. En concreto, se introduce una clase genérica de certificados digitales en X-Road, reduciendo la dependencia de perfiles por país. Además, se describe la evolución de MARISMA para entornos públicos, incorporando un meta-patrón de riesgos fundamentado en CAPEC y la matriz MITRE ATT&CK. Finalmente, se subraya la relevancia de la Gobernanza Estratégica para unificar políticas y procesos de protección conforme a los objetivos institucionales, logrando una protección más robusta y adaptable para los servicios digitales de Gobierno.**

***Index Terms*—Ciberseguridad, e-Gobierno, Framework holístico, X-Road, MARISMA, GAUCHO**

Tipo de contribución: *Investigación original*

I. Introducción

La creciente digitalización de los servicios gubernamentales ha puesto de manifiesto nuevos desafíos en materia de seguridad de la información, motivando la búsqueda de soluciones integrales que aborden tanto la protección de datos como la interoperabilidad de los sistemas [1]. Según Agawu [2], la evolución del e-Gobierno pasa por servicios más 'inteligentes' y ubicuos, sin descuidar la protección de datos y la confianza de los usuarios. En ese sentido, en las administraciones públicas, la heterogeneidad de plataformas y la necesidad de intercambiar información sensible refuerzan la importancia de adoptar un enfoque holístico de ciberseguridad, donde la gestión de riesgos, la gobernanza y la infraestructura tecnológica trabajen de forma coordinada [3].

En este contexto surge GAUCHO *Government Adaptive Unified Cybersecurity Holistic Orchestration* [4], un framework de ciberseguridad para e-Gobierno que integra tres pilares fundamentales:

- Gestión de riesgos con MARISMA. Este componente se basa en la sistematización de activos y vulnerabilidades, así como en la aplicación de controles adaptativos. De acuerdo a Sánchez et al. [5], MARISMA aporta una metodología ágil para la identificación, evaluación y tratamiento de riesgos; está alineada con estándares internacionales y en su versión extendida para administraciones públicas se apoya en catálogos de amenazas como CAPEC y la matriz MITRE ATT&CK.
- Interoperabilidad segura con X-Road. Tal y como plantea Adeodato et al. [6], X-Road ofrece un mecanismo de intercambio de datos fiable entre los distintos organismos gubernamentales, garantizando la confidencialidad, integridad y disponibilidad de la información. Entre sus ventajas destaca la posibilidad de centralizar la validación de certificados y el control de accesos, si bien la versión estándar requiere personalizaciones por país que dificultan la escalabilidad.
- Gobernanza estratégica de ciberseguridad. En los trabajos previos de investigación que han dado lugar al framework GAUCHO, se llevó a cabo una revisión sistemática que permitió identificar un conjunto de carencias en los trabajos existentes orientados a la ciberseguridad para e-Gobierno. A partir de este hallazgo, se han definido 16 características deseables que son cubiertas por cada uno de los componentes que integran el nuevo framework. En ese sentido, el componente de gobernanza estratégica de la ciberseguridad, tiene como objetivo dar cumplimiento a aquellas características enfocadas en la coordinación eficiente de los procesos de protección y su alineación con los objetivos institucionales; GAUCHO propone un módulo de gobernanza que defina políticas, roles y métricas de evaluación continua. Este módulo promueve la toma de decisiones basada en riesgos y la evolución del marco de seguridad en función de las necesidades de cada entidad pública, teniendo en cuenta la premisa de Zhang et al. [7], en cuyo trabajo subrayan que si solo nos enfocamos en la infraestructura o en la parte legal sin integrar la dimensión social (capacitación, cultura de ciberseguridad, percepción ciudadana), los resultados no serán óptimos.

La conjunción de estos tres ejes —gestión adaptativa de riesgos, interoperabilidad segura y gobernanza estratégica de la ciberseguridad — permite abordar los requerimientos específicos del e-Gobierno de manera transversal. A lo largo de este trabajo, se profundiza en la ampliación de X-Road mediante una nueva clase genérica de certificados digitales y en la adaptación de MARISMA al Sector Público, que intro-

duce un meta-patrón de riesgos basado en CAPEC y MITRE ATT&CK. Dichos aportes se combinan con las directrices de gobernanza para conformar una arquitectura de ciberseguridad integral más flexible, escalable y enfocada en la protección efectiva de servicios electrónicos gubernamentales.

El resto del artículo se organiza como sigue: en la Sección II se revisan los trabajos relacionados que sustentan esta propuesta. La Sección III describe el diseño y la estructura de GAUCHO, centrándose en la clase genérica de certificados para X-Road y la extensión de MARISMA. Posteriormente, en la Sección IV se exponen los resultados preliminares de validación, mientras que la Sección V recoge las conclusiones y las posibles líneas de trabajo futuro.

II. Trabajos relacionados

La ciberseguridad en e-Gobierno se ha abordado desde múltiples ángulos, integrando metodologías de análisis de riesgos, marcos de gobernanza y soluciones tecnológicas específicas. A continuación se revisan algunos trabajos afines que contextualizan la propuesta de GAUCHO:

- Frameworks de ciberseguridad para e-Gobierno. Diversas instituciones y organismos internacionales —como ISO (serie 27000), ITU, NIST o ENISA— han publicado lineamientos y recomendaciones para fortalecer la seguridad de servicios gubernamentales [8]. No obstante, muchos de estos enfoques suelen centrarse en aspectos técnicos o en el cumplimiento normativo, sin integrar de forma integral la interoperabilidad y la gestión adaptativa de riesgos necesarias en entornos públicos complejos [9]. En este sentido, el mercado ha presenciado la aparición de propuestas académicas de corte más holístico, que recomiendan unificar la perspectiva de gobernanza, los controles técnicos y la continua evaluación del riesgo [10].
- Gestión adaptativa de Riesgos. El rol de la gestión de riesgos resulta esencial para hacer frente a las amenazas en el ámbito gubernamental. Stone [11] subraya que, en este contexto, la adopción de metodologías formales de análisis y tratamiento de riesgos reduce de forma significativa la exposición a ciberataques. En los últimos años, se han desarrollado iniciativas específicas que adaptan estándares de riesgo a la realidad de los organismos públicos. Por ejemplo, algunas propuestas basadas en ISO/IEC 27005 o en la metodología MAGERIT han puesto el acento en el catalogado de activos de información sensibles y la medición de probabilidades de incidentes [12]. No obstante, persiste la ausencia de taxonomías suficientemente dinámicas que permitan incorporar de forma ágil nuevos escenarios de ataque. En este panorama, MARISMA cobra relevancia al introducir un enfoque basado en meta-patrones, que facilita la adaptación a diferentes entornos de riesgo. Ejemplo de ello son MARISMA-BiDa [13], orientado a big data, MARISMA-CPS [14] para entornos ciberfísicos y MARISMA-BP [15] centrado en procesos de negocio. Gracias a esta versatilidad, GAUCHO incorpora catálogos de amenazas basados en CAPEC y la matriz MITRE ATT&CK, contribuyendo así a una mayor precisión en la identificación y tratamiento de los riesgos.
- Interoperabilidad y X-Road. El concepto de X-Road surgió con el objetivo de garantizar un intercambio seguro de datos entre entidades gubernamentales, facilitando la interoperabilidad y la trazabilidad [16]. Algunos países —caso de Estonia— han consolidado su infraestructura digital sobre esta plataforma, generando experiencias pioneras en e-Governance. Sin embargo, su adopción global implica desafíos, entre los que destaca la necesidad de perfiles de certificados específicos por país, encareciendo la personalización y dificultando la escalabilidad en contextos multinacionales. Como parte del enfoque holístico de GAUCHO, se ha propuesto la creación de una clase genérica de certificados para X-Road, abordando dicha limitación y mejorando la flexibilidad en su despliegue.
- Gobernanza estratégica de la seguridad. Gran parte de la literatura coincide en que la mera aplicación de controles técnicos resulta insuficiente sin una gobernanza sólida que defina responsabilidades, procesos de supervisión y métricas de mejora continua [17]. Se proponen modelos de ciberseguridad centrados en el alineamiento con los objetivos organizacionales —por ejemplo, COBIT o marcos híbridos inspirados en ITIL—, pero pocos de ellos se focalizan directamente en las particularidades de la Administración Pública y en la coordinación de múltiples entidades [18]. *GAUCHO* integra un módulo de Gobernanza Estratégica que, además de establecer políticas y roles, promueve la integración de la gestión de riesgos (MARISMA) y de la infraestructura de interoperabilidad (X-Road) bajo criterios unificados.

La aceptación ciudadana de los servicios electrónicos gubernamentales se ve fuertemente influida por la facilidad de uso y la confianza que los usuarios depositan en las plataformas. Según Carter et al. [19], aspectos como la disposición a confiar en la tecnología y la percepción de utilidad tienen un impacto determinante en la adopción efectiva del e-Gobierno. Este hallazgo refuerza la idea de que un marco holístico de ciberseguridad —como el que se propone en GAUCHO— debe tener en cuenta no solo la interoperabilidad y la gestión adaptativa de riesgos, sino también la experiencia y percepción de los ciudadanos, como usuarios de los servidos digitales del e-Gobierno.

Cabe destacar que un alto nivel de adopción de e-Gobierno no necesariamente garantiza mayor seguridad. Según Thompson et al. [20], en su análisis comparativo de Australia y Tailandia, la mera popularidad de los servicios digitales gubernamentales no se traduce siempre en infraestructuras ciberseguras sólidas. Este punto subraya la necesidad de frameworks integrales —como GAUCHO— que aborden tanto la interoperabilidad (X-Road) como la gestión adaptativa de riesgos (MARISMA) y la gobernanza estratégica.

En el ámbito gubernamental, la ciberseguridad enfrenta retos particulares derivados de limitaciones presupuestarias y de personal especializado. Según Hossain et al. [21], en su revisión sistemática sobre ciberseguridad en gobiernos, existe una carencia de marcos unificados que guíen la protección de datos e infraestructuras críticas en este nivel, al igual que manifiesta Lino [22] tras analizar varios incidentes en agencias de gobierno estadounidenses. Nuestra propuesta de extensión

de MARISMA para e-Gobierno busca precisamente solventar esas limitaciones al brindar una metodología adaptativa y escalable para riesgos en estos entornos.

En suma, la literatura refleja la necesidad de contar con soluciones holísticas que abarquen la gobernanza, la gestión flexible de riesgos y la interoperabilidad segura. Los avances en X-Road y en metodologías de riesgo apuntan en esa dirección, pero suelen afrontarse de modo aislado. La contribución de GAUCHO radica en unir dichos componentes en un solo framework y, de forma particular, extender tanto X-Road como MARISMA para satisfacer los requerimientos de la administración pública, tal como se profundizará en la siguiente sección.

III. Propuesta y Diseño

La propuesta del framework GAUCHO integra la interoperabilidad segura ofrecida por X-Road con un enfoque de gestión adaptativa de riesgos, basado en MARISMA, y un modelo de Gobernanza Estratégica de la Ciberseguridad, con el objetivo de ofrecer una solución holística para entornos de e-Gobierno. A continuación, se describe la arquitectura conceptual de GAUCHO, detallando el rol de cada uno de sus componentes, sus aportes técnicos específicos en este trabajo y la forma en que interactúan de manera coordinada. Posteriormente, se desarrolla el eje central de esta contribución, centrado en la implementación concreta de dos mejoras clave:

- una clase genérica de perfiles de certificados digitales para X-Road,
- y un nuevo meta-patrón de riesgos integrable en MARISMA, específicamente adaptado a entornos gubernamentales.

III-A. Visión General de la Arquitectura GAUCHO

El objetivo principal de GAUCHO es orquestar, de manera coherente, los procesos de seguridad que surgen en la administración pública digital, unificando la interoperabilidad (para compartir datos de forma confiable), la gestión de riesgos (para identificar y mitigar amenazas a los activos estatales) y la gobernanza estratégica (para asegurar que las decisiones de seguridad estén alineadas con los objetivos institucionales).

III-A1. Interoperabilidad Segura con X-Road: X-Road permite el intercambio seguro de datos entre entidades a través de mecanismos criptográficos, certificados digitales y autenticación mutua. Cada miembro de la red está identificado por atributos únicos, incluidos en los certificados utilizados para firmar y asegurar las transacciones. La interoperabilidad multinacional exige desacoplar estos certificados de estructuras fijas por país, lo cual ha motivado el desarrollo de una clase genérica de perfiles, detallada en la sección III-B.

III-A2. MARISMA: Gestión Adaptativa de Riesgos: MARISMA es un marco metodológico orientado a la evaluación y gestión de riesgos, diseñado para ajustarse a distintos entornos mediante la incorporación de patrones de amenazas específicos. En el contexto de GAUCHO, MARISMA actúa como el componente de análisis continuo de ciberseguridad, con capacidades ampliadas para entornos gubernamentales. La integración concreta de catálogos como CAPEC y la matriz MITRE ATT&CK se detalla en la sección III-C.

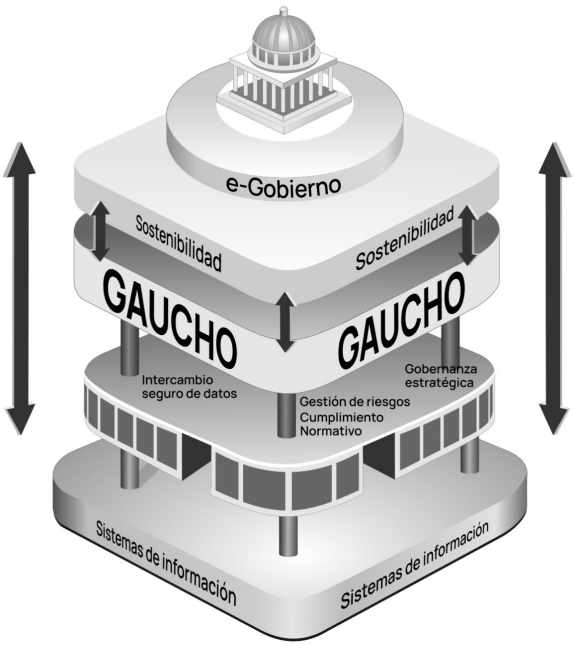


Figura 1. Arquitectura integrada del framework GAUCHO.

III-A3. Gobernanza Estratégica de la Seguridad: Define políticas, responsabilidades y procesos necesarios para garantizar que se cumplan con todas las características deseables identificadas en la revisión sistemática. Contempla además la definición de métricas de gobierno (p.ej., KPIs y KRIs) que posibilitan medir la eficacia de las acciones de ciberseguridad a lo largo del tiempo [24].

III-A4. Sostenibilidad del marco: Un aspecto clave de la arquitectura GAUCHO es su sostenibilidad operativa, entendida como la capacidad del marco para adaptarse a distintos contextos institucionales y jurisdiccionales sin requerir modificaciones estructurales o de bajo nivel. Por ejemplo, la clase genérica de certificados desarrollada en este trabajo permite a cada Gobierno utilizar su propio código de país sin recompilar el núcleo de X-Road, lo cual simplifica la adopción y reduce costos de mantenimiento. De igual modo, la extensión de MARISMA mediante meta-patrones permite actualizar el catálogo de amenazas sin redefinir la lógica base del sistema. Este enfoque modular y desacoplado facilita la evolución del marco ante nuevas normativas, tecnologías o amenazas, asegurando su vigencia en el tiempo con un mínimo esfuerzo de reconfiguración técnica o normativa.

De esta forma, GAUCHO busca que toda integración de servicios electrónicos gubernamentales (vinculada a X-Road) esté monitorizada y gestionada desde una única visión de la ciberseguridad (vía MARISMA) y bajo una dirección estratégica común (gobernanza), lo cual abarca tanto la capa tecnológica como la organizativa. En la Figura 1 se ilustra cómo el extremo superior se orienta hacia el ciudadano, articulando servicios de e-Gobierno; mientras que el extremo inferior representa los sistemas de información internos —el activo principal a proteger— sobre los que actúan las medidas integrales de seguridad y gobernanza definidas por GAUCHO.

Un rasgo distintivo del enfoque GAUCHO es la existencia de mecanismos de retroalimentación operativa entre sus

componentes, integrados además en un contexto de ciberseguridad más amplio. La interoperabilidad habilitada por X-Road permite exponer y consumir servicios electrónicos entre organismos, pero su operación está condicionada por las políticas establecidas desde la gobernanza estratégica y por la evaluación dinámica de riesgos realizada por MARISMA. Este último se nutre tanto de catálogos estructurados como CAPEC y MITRE ATT&CK, como de señales provenientes de fuentes técnicas externas, tales como sistemas de correlación de eventos, cortafuegos, sistemas de detección y prevención de intrusiones o, incluso de soluciones de inteligencia de amenazas.

Por ejemplo, ante la detección de un incidente por parte del SIEM, se puede disparar una alerta que activa un nuevo ciclo de evaluación en MARISMA; o bien, la incorporación de un nuevo patrón de ataque en CAPEC puede provocar la recalibración automática de niveles de riesgo asociados a servicios o activos específicos. Si se identifica una vulnerabilidad crítica en un componente de infraestructura (como X-Road), MARISMA puede proponer medidas de mitigación, que incluyen ajustes de confianza, restricción de acceso, o modificaciones en la exposición del servicio. A partir de esa evaluación, el componente de Gobernanza Estratégica toma decisiones conforme a las políticas institucionales, pudiendo intervenir sobre flujos de X-Road o definir campañas de concienciación, procedimientos de respuesta a incidentes o revisiones de cumplimiento.

Esta lógica de *detección–evaluación–decisión–acción* posiciona a GAUCHO como un marco orquestador capaz de coordinar soluciones técnicas, procesos operativos y directrices normativas, garantizando una respuesta adaptativa ante amenazas emergentes sin desarticular la operación de los servicios digitales del Estado.

III-B. Ampliación de X-Road: Creación de una Clase Genérica de Perfiles de Certificados

A continuación, se describe la necesidad de adaptar X-Road mediante la introducción de una clase genérica de certificados. Primero, se exponen los antecedentes y la forma en que X-Road gestiona los perfiles de certificado; después, se presenta la nueva clase propuesta, explicando su diseño y las ventajas que ofrece al liberarse del country code fijo.

III-B1. Antecedentes: ¿Qué es un Perfil de Certificado en X-Road?: Según la documentación oficial [25], un perfil de certificado define la manera en que se almacenan, en el propio certificado X.509, atributos necesarios para que X-Road funcione (por ejemplo, instanceIdentifier, memberClass, memberCode). El problema reside en que muchos de estos perfiles son específicos de un país, con un código de país (p.ej., “EE”) predifinido en el código fuente. Al incluirlos en el campo C (country) o en el serialNumber, se limitaba la adopción multinacional de X-Road.

Por ejemplo, las implementaciones para Finlandia incorporan la clase java FiVRKCertificateProfileInfoProvider y otras (p.ej. FiVRKAuthCertificateProfileInfo o FiVRKSignCertificateProfileInfo) donde el código de país se fija en “FI” y el serialNumber se compone de [instanceIdentifier]/[serverCode]/[memberClass]. Esta configuración obliga a recompilar el software si otro país desea usar una codificación distinta.

III-B2. Clase Genérica de Certificados: Para superar esta limitación, en el marco de GAUCHO se ha desarrollado una clase genérica de certificado que:

- Omite la codificación fija del país, permitiendo que el campo C se reciba como parámetro desde la interfaz de usuario al momento de generar la solicitud de firma de certificado (CSR) desde el Servidor de Seguridad.
- Mantiene la obligación de incluir los datos requeridos por X-Road en el serialNumber (p.ej. [instanceIdentifier]/[memberClass]/[memberCode]), pero sin referencias a un país concreto.
- Facilita la configuración en el Servidor Central (Central Server), al poder seleccionar un perfil de certificados genérico en lugar de un perfil de certificado específico por país.

Para ello, se implementa una variante análoga a la mostrada en la familia FiVRK*, pero parametrizable. Las clases fundamentales de la propuesta serían:

- GenericCertificateProfileInfoProvider: clase principal que implementa la interfaz CertificateProfileInfoProvider, retornando instancias específicas para la autenticación (GenericAuthCertificateProfileInfo) y la firma (GenericSignCertificateProfileInfo).
- GenericSignCertificateProfileInfo: gestiona la creación/lectura del serialNumber que contendrá [instanceIdentifier]/[memberCode]/[memberClass].
- GenericAuthCertificateProfileInfo: define los campos de autenticación (en X-Road, la información del certificado de autenticación no se utiliza activamente, pero sirve para coherencia y auditoría).
- GenericSubjectClientIdDecoder: encargado de parsear el serialNumber en los atributos que X-Road requiere.

A continuación, en la Figura 2 se presenta un diagrama de clases que ilustra la extensión genérica de X-Road para gestionar certificados sin depender de un country code fijo. Como se observa, GenericCertificateProfileInfoProvider implementa la interfaz básica (CertificateProfileInfoProvider) y define métodos como getCSRCountryCode(...) para recibir de forma flexible el valor del campo C (por ejemplo, “AR”, “ES”). A su vez, este proveedor retorna instancias de GenericAuthCertificateProfileInfo y GenericSignCertificateProfileInfo, cada una encargada de crear el Subject Distinguished Name (DN) y analizar el certificado según sea para autenticación o firma. De esta manera, el serialNumber incluirá [instanceIdentifier]/[memberClass]/[memberCode] sin necesidad de codificar un país en el core de X-Road. La clase genérica potencia así la escalabilidad y la interoperabilidad, dado que cada jurisdicción puede inyectar su propio country code, alineando esta solución con el enfoque holístico de GAUCHO.

En el esquema anterior, la clase GenericCertificateProfileInfoProvider recibe de forma dinámica el valor de countryCode mediante el método getCSRCountryCode(), y delega a GenericAuthCertificateProfileInfo o GenericSignCertificateProfileInfo la construcción y el análisis de los certificados, respetando la lógica de X-Road sin anclar la configuración a un país.

III-B3. Beneficios de la Clase Genérica: La solución propuesta en este trabajo se centra en el desarrollo de una nueva clase de perfiles de certificados digitales genéricos para X-Road. Como se mencionó en la sección anterior, los perfiles

de certificados existentes son mayoritariamente específicos de cada país, lo que genera limitaciones en la interoperabilidad y escalabilidad de la plataforma. Para abordar estas deficiencias, se propone una clase de perfil que no contenga un código de país codificado o predefinido por código fuente, lo que le permitirá ser utilizada en cualquier contexto nacional o internacional.

Esta aportación supone un conjunto de beneficios:

- Escalabilidad: Ya no es necesario recompilar X-Road al cambiar de país o de requisitos locales.
- Menor complejidad: Se configura todo en el Servidor Central, reduciendo el overhead de crear un nuevo proyecto de certificado en el repositorio oficial de X-Road.
- Facilidad de integración: Permite la conexión de organismos que operan bajo diferentes normativas, sin duplicar esfuerzos en codificación de perfiles particulares.
- Alineación con la visión holística: Al integrarse con MARISMA y la gobernanza, los certificados no solo garantizan la interoperabilidad, sino que también se adaptan a la gestión continua de riesgos y a las políticas de seguridad definidas en GAUCHO.

La siguiente tabla comparativa, Tabla I, ilustra de forma resumida los aportes de esta contribución.

Tabla I
TABLA COMPARATIVA DE PERFILES DE CERTIFICADOS EN X-ROAD

Criterio	Clases Predefinidas	Clase Genérica
Configuración de país	NO	SI
Edición de Código	SI	NO
Escalabilidad	NO	SI
Interoperabilidad	SI	SI

III-C. Extensión de MARISMA para Gobierno: Nuevo meta-patrón de Riesgos

Tal y como se apuntó en la Sección anterior, MARISMA aporta una metodología de riesgo dinámica que, en su versión para e-Gobierno, se extiende con el soporte a catálogos CAPEC y MITRE ATT&CK. El objetivo es hacer frente a ataques dirigidos específicamente a infraestructuras gubernamentales,

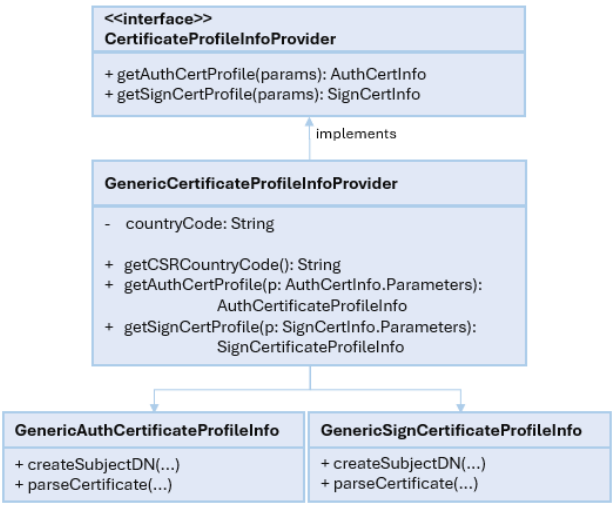


Figura 2. X-Road. Diagrama de Clases Perfil de certificados genéricos

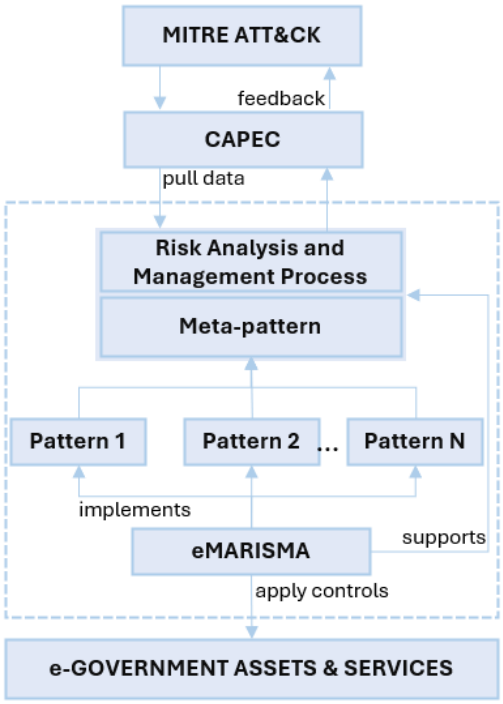


Figura 3. Esquema general del meta-patrón riesgos de MARISMA para e-Gobierno

tales como robo de credenciales ciudadanas, manipulación de bases de datos, interrupción de servicios críticos, entre otros, a partir del siguiente enfoque:

- Metodología: Se definen patrones de riesgo sectoriales que se mapearán a las técnicas de ataque documentadas en ATT&CK.
- Actualización continua: Permite sincronizar de forma regular el catálogo de amenazas, de modo que se puedan añadir nuevas entradas o tácticas detectadas globalmente.
- Integración con políticas de defensa: Al saber qué formas de ataques son más probables, se actualizan de manera dinámica las políticas de protección implementadas mediante las distintas soluciones existentes en la infraestructura tecnológica gubernamental.

En la Figura 3 se presenta un esquema general del meta-patrón implementado en MARISMA, adonde se plasma la naturaleza cíclica de la gestión de riesgos: MARISMA extrae conocimiento de repositorios, aplica esa información para proteger servicios gubernamentales, y a su vez puede generar una retroalimentación que enriquezca la base de conocimiento global, especialmente cuando se detectan ataques novedosos. Así, las flechas bidireccionales reflejan un ecosistema interactivo donde no solo se consume información de CAPEC y MITRE, sino que también se puede colaborar con ellas (si el escenario o la política de la entidad pública lo permite).

III-D. Módulo de Gobernanza Estratégica de Seguridad

Finalmente, el tercer componente de GAUCHO actúa como eje integrador entre los mecanismos técnicos y las decisiones institucionales. La gobernanza estratégica permite definir políticas de seguridad, gestionar respuestas a incidentes, establecer mecanismos de auditoría, y coordinar acciones

entre múltiples actores. Su implementación no responde a una herramienta específica, sino a una lógica de supervisión que recibe información proveniente de MARISMA y de capas técnicas externas, y ejecuta acciones correctivas en función de criterios definidos por la organización.

La gobernanza puede ordenar el retiro de servicios en X-Road, la revocación de certificados, la activación de campañas de concienciación o la revisión de acuerdos interinstitucionales. Esta capacidad de actuar en múltiples dimensiones permite que el marco GAUCHO opere de forma coherente ante incidentes, priorizando la continuidad de los servicios esenciales y el cumplimiento normativo.

IV. Validación y Resultados Preliminares

Para evaluar la factibilidad y los beneficios de la propuesta, se han llevado a cabo pruebas preliminares centradas en dos aspectos clave: (1) la implementación de la nueva clase genérica de certificados digitales en X-Road, y (2) la eficacia de la extensión de MARISMA para e-Gobierno apoyada en CAPEC y MITRE ATT&CK. A continuación se describen los escenarios de prueba y los hallazgos iniciales.

IV-A. Escenario de Prueba: Despliegue Multijurisdiccional de X-Road

En esta prueba se simula la adopción de X-Road por parte de distintos gobiernos, cada uno con su propia codificación de país y su propio entorno de confianza, verificando que la clase genérica de certificados (GenericCertificateProfile) elimina la necesidad de modificar el código fuente de X-Road para cada país. El objetivo es demostrar que varias instancias gubernamentales pueden interoperar sin fricciones ni recompilaciones adicionales.

Objetivos.

- Validar que cada gobierno (o jurisdicción) pueda configurar su propio country code (C) en los certificados o incluso omitirlo, manteniendo la coherencia con los campos requeridos (instanceIdentifier, memberClass, memberCode) en el serialNumber.
- Confirmar que la infraestructura de X-Road (Central Server, Trust Services y Security Servers) se integre correctamente en un escenario con múltiples autoridades y parámetros nacionales diferentes.

Gobiernos Participantes. Se definen dos entornos: Gobierno A y Gobierno B. Cada uno elige un country code distinto: “AR”, “CL”.

Instancia de X-Road en Cada Gobierno. Cada gobierno despliega su propio Central Server (CS) y un conjunto de Security Servers (SS) para los ministerios o entidades participantes. Emplean la clase GenericCertificateProfileInfoProvider para generar los CSR de firma, que permitirá a los SS especificar por parámetro el valor del campo C (country) deseado.

Trust Services. Complementariamente, cada gobierno agrega un Trust Service que proporcione las autoridades certificadoras (CA), servicios de validación de certificados -OCSP- y los servicios de time stamping, de modo que cada uno de ellos pueda validar sus certificados e implementar otros servicios de seguridad consumidos por X-Road.

Registro e Intercambio. El Gobierno A registra su instancia y la de sus organismos en el CS “A”. Lo mismo hace el Gobiernos B en sus respectivos entornos, sin necesidad de recompilar X-Road para incluir sus códigos de país correspondientes.

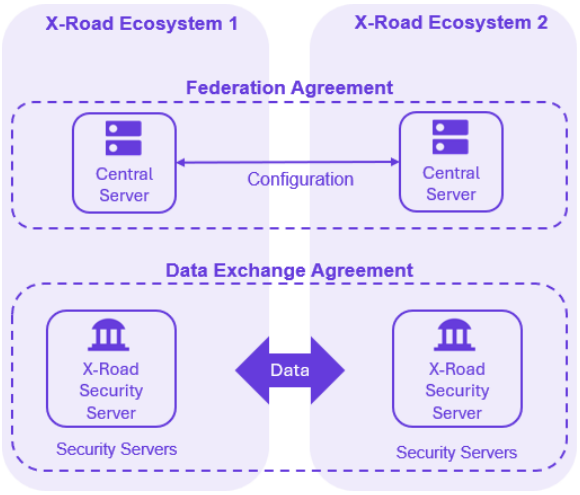


Figura 4. X-Road. Federación de instancias inter-jurisdiccionales.

Cada infraestructura asume que los serialNumber de los certificados contienen [instanceIdentifier]/[memberClass]/[memberCode].

Interoperabilidad. Para comprobar la comunicación inter-jurisdiccional, se habilita un intercambio de servicios bajo el concepto de Federación de Ecosistemas: el Gobierno A llama a un servicio alojado en el Gobierno B, y viceversa. Se comprueba que la autenticación mutua (TLS) y la validación de certificados se ejecutan con éxito.

Resultados esperados. Cada Gobierno personaliza su country code “C” sin editar ni recompilar el núcleo de X-Road. Los Central Servers aceptan los CSR generados con la clase genérica de certificados y reconocen adecuadamente los atributos instanceIdentifier, memberClass y memberCode en el serialNumber. Los mensajes fluyen con autenticación y trazabilidad, confirmando la validez del nuevo enfoque multinacional. En la Figura 4 se ilustra este escenario a través de un esquema con dos instancias de X-Road (cada una con sus propios componentes), que interoperan bajo el concepto de Federación, sin requerir la recompilación del código base para incorporar perfiles específicos para cada país.

Hallazgos principales.

- Escalabilidad: Se confirma que cada jurisdicción puede sumarse a la federación sin pasos adicionales en la base de código X-Road.
- Flexibilidad: El country code en el certificado se vuelve un parámetro configurable, adaptándose a la normativa local de cada Gobierno.
- Ahorro de costes: Menor dependencia de perfiles específicos por país; se evita crear variantes como FiVRK, EE, etc., para cada región.
- Mayor cooperación: El escenario demuestra que con un reconocimiento mutuo de CAs, los Gobiernos A y B intercambian información de manera segura y armoniosa.

IV-B. Validación de la Extensión de MARISMA para e-Gobierno

Se diseñó una prueba de concepto centrada en la detección y priorización de amenazas que afectan a servicios de e-Gobierno.

Catálogos incorporados. Se implementó un script que sincroniza periódicamente la base de amenazas de MARISMA con la última versión de CAPEC (ataques catalogados #1, #2, #3, etc.) y con la matriz ATT&CK (tácticas Txxxx), obteniendo a la vez las contramedidas aplicables. Se definió un *meta-patrón para e-Gobierno* que agrupa amenazas relevantes para servicios gubernamentales y se mapean a sus activos específicos.

Metodología de evaluación. Se seleccionaron cinco activos críticos: un sistema de expedientes digitales, un portal de trámites para ciudadanos, un portal educativo, una base de datos de registros de personas y una plataforma de información hospitalaria. MARISMA realizó una evaluación de riesgos inicial (probabilidad e impacto), añadiendo 12 nuevos vectores de ataque provenientes de la última actualización de CAPEC/ATT&CK. Se definieron controles prioritarios (e.g., limitaciones de acceso, monitoreo de logs, pruebas de penetración periódicas).

Resultados. Se registraron 15 riesgos nuevos no contemplados en el catálogo base de MARISMA, relacionados con técnicas de escalada de privilegios y phishing dirigido a gestores de los servicios. El análisis automático propuso contramedidas tales como endurecer políticas de autenticación multifactor, revisar los protocolos de cifrado en redes privadas virtuales y sensibilizar a los empleados. El 73 % de los riesgos recién incorporados recibió un tratamiento de *Alta Prioridad*, elevando el nivel de concienciación sobre amenazas emergentes.

En la Tabla II se muestra de forma concisa cómo la combinación de CAPEC y MITRE ATT&CK permite a MARISMA identificar y categorizar amenazas concretas (columna "CAPEC / ATT&CK"), asignarles un nivel de probabilidad, impacto y riesgo (columnas siguientes) y, finalmente, recomendar contramedidas específicas basadas en los IDs de mitigación de ATT&CK (última columna). Gracias a este formato, la administración pública puede priorizar mejor los recursos de protección y desplegar controles eficaces contra las técnicas de ataque más relevantes.

Tabla II
EXTRACTO DE RIESGOS EMERGENTES AL INTEGRAR CAPEC Y MITRE ATT&CK EN MARISMA

CAPEC / ATT&CK	Probabilidad	Impacto	Riesgo	Contramedida
CAPEC-93 / T1059	M	A	A	M1040, M1036
CAPEC-463 / T1189	A	M	A	M1026
CAPEC-27 / T1078	A	C	C	M1052, M1057
CAPEC-38 / T1566	M	M	M	M1032
CAPEC-10 / T1569	M	A	A	M1036, M1042
CAPEC-21 / T1210	B	A	M	M1021
CAPEC-660 / T1499	B	M	B	M1003, M1031
CAPEC-48 / T1190	A	C	C	M1048
CAPEC-39 / T1592	M	M	M	M1057
CAPEC-112 / T1053	A	A	A	M1038, M1055

C:Crítico; **A**:Alto/a; **M**:Medio/a; **B**:Bajo/a

IV-C. Discusión de los Resultados

A continuación, se analizan los principales hallazgos y su relevancia para la adopción de GAUCHO en entornos gubernamentales. En primer lugar, se discute la contribución de la clase genérica de certificados en X-Road; a continuación, se revisa el rol del meta-patrón de MARISMA para e-Gobierno y, finalmente, se exponen las limitaciones encontradas y las perspectivas de evolución futura.

IV-C1. Sobre la Clase Genérica de Certificados: La prueba demostró que es posible implementar múltiples instancias de X-Road con distintos códigos de país sin tocar la lógica interna de X-Road, confirmando la viabilidad de la propuesta. La simplificación en la generación de CSR y la configuración del Servidor Central disminuye la curva de aprendizaje y evita la recompilación. Se alinea con la visión holística al integrar una solución adaptable con la capa de gobernanza y la monitorización de riesgos.

IV-C2. Sobre el Meta-patrón de MARISMA para e-Gobierno: El uso de CAPEC y ATT&CK enriquece la detección y categorización de amenazas, algo esencial en un entorno donde los ataques evolucionan rápidamente. La clasificación de riesgos resultante permite focalizar esfuerzos de mitigación allí donde el impacto potencial es mayor, reflejando la adaptabilidad de MARISMA. Al unificar estos hallazgos con el módulo de gobernanza, se sientan las bases para una toma de decisiones proactiva, apuntalada por información actualizada sobre amenazas reales.

IV-C3. Limitaciones y Perspectiva Futura: Este escenario de validación no representa todas las complejidades posibles de un entorno gubernamental real (por ejemplo, integra solo dos organismos, falta un entorno de producción con miles de transacciones diarias). Se prevé extender el alcance de la integración hacia la federación entre distintas instancias de X-Road (varios países o gobiernos locales), así como profundizar en la automatización de controles de seguridad sugeridos por MARISMA. En conjunto, los resultados preliminares indican que la propuesta viabiliza la adopción multi-jurisdiccional de X-Road, a la vez que mejora la capacidad de las administraciones para identificar, priorizar y abordar riesgos emergentes. La siguiente sección presentará las conclusiones y perspectivas de evolución del framework GAUCHO.

V. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo ha presentado extensiones técnicas sobre dos componentes del framework GAUCHO: una clase genérica de perfiles de certificados digitales para X-Road, que facilita la interoperabilidad multinacional sin modificar el núcleo de la plataforma, y un meta-patrón de riesgos integrable en MARISMA, basado en CAPEC y MITRE ATT&CK, diseñado para entornos gubernamentales. Ambas contribuciones fortalecen la capacidad de adaptación del marco ante amenazas emergentes y refuerzan su aplicabilidad en contextos públicos reales.

No obstante, la propuesta aún tiene márgenes de evolución. Se planea escalar la integración entre X-Road y MARISMA a escenarios productivos con mayor volumen transaccional y más organismos participantes, evaluando así el comportamiento a gran escala y la latencia en intercambios críticos. También se prevé profundizar en la federación de instancias X-Road entre diferentes países y explorar la automatización avanzada de controles, aprovechando inteligencia de amenazas en tiempo real.

En resumen, los resultados reflejan que la confluencia de interoperabilidad segura, gestión dinámica de riesgos y gobernanza estratégica de la ciberseguridad posibilita un entorno de e-Gobierno más resiliente y confiable. Con este enfoque, GAUCHO establece las bases para un modelo integral de ciberprotección que sea adaptable a distintas realidades jurídicas y amenazas globales, impulsando la colaboración y la confianza entre ciudadanos, agencias gubernamentales y actores internacionales.

VI. Agradecimientos

Este trabajo se ha desarrollado gracias a la financiación proporcionada por los siguientes proyectos: Proyecto Di4SPDS (CHIST-ERA grant- PCI2023145980-2) financiado por MCIN/AEI y cofinanciado por la Unión Europea, KOSMOS-UCLM (PID2024-155363OB-C44) financiado por MCIN/AEI/10.13039/501100011033; ALBA-UCLM (TED2021-130355B-C31) financiado por CIN/AEI/10.13039/501100011033/ Unión Europea NextGenerationEU/PRTR, y MESIAS (2022-GRIN-34202) financiado por FEDER.

Referencias

[1] Willemson Jan, Ansper Arne: .A secure and scalable infrastructure for inter-organizational data exchange and e-Government applications", en *2008 Third International Conference on Availability, Reliability and Security*, 2008.

[2] Agawu, Emefa: "What's Next for e-Government: Innovations in e-Government through a Cybersecurity Lens Symposium", en *Neb. L. Rev.*, vol. 96, n. 2, pp. 364–383, 2017.

[3] More Manda, Judy Backhouse: .Addressing trust, security and privacy concerns in e-Government integration, interoperability and information sharing through policy: a case of South Africa", en *CONF-IRM 2016 Proceedings*, 2016.

[4] Figueroa, Víctor, Sánchez, Luis E., Santos-Olmo, Antonio, Rosado, David G., Fernández-Medina, Eduardo: "Towards a Holistic Cybersecurity Framework for e-Governments", en *2024 IEEE Biennial Congress of Argentina (ARGENCON)*, pp. 18, 2024.

[5] Luis E. Sánchez, Antonio Santos-Olmo, David G. Rosado, Carlos Blanco, Manuel A. Serrano, Haralambos Mouratidis, Eduardo Fernández-Medina: "MARISMA: A modern and context-aware framework for assessing and managing information cybersecurity risks", en *Computer Standards & Interfaces*, vol. 92, pp. 103935, 2025.

[6] Adeodato, Rodrigo, Pournouri, Sina: "Secure implementation of e-governance: A case study about estonia", en *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, pp. 397–429, 2020.

[7] Zhang Huiping, Tang Zhiwei, Jayakar Krishna: .A socio-technical analysis of China's cybersecurity policy: Towards delivering trusted e-Government services", en *Telecomm. Policy*, vol. 42, n. 5, pp. 409–420, 2018.

[8] Mishra Shailendra, Alowaidi Majed A, Sharma Sunil Kumar: Ïmpact of security standards and policies on the credibility of e-Government", en *J. Ambient Intell. Humaniz. Comput.*, 2021.

[9] Mayayise, Thembekile O: .A security-based e-Government adoption framework: An outcome of a systematic literature review", en *Proceedings of Eighth International Congress on Information and Communication Technology*, pp. 265–282, 2024.

[10] Santos-Olmo, A., Sánchez, L.E., Rosado, D.G., Fernández-Medina, E., Mouratidis, H., Serrano, M.A., Blanco, C.: "Towards an integrated risk analysis security framework according to a systematic analysis of existing proposals.", en *Frontiers of Computer Science*, vol. 18, n. 3, 2024.

[11] Alexander Gregory Stone: "The role risk-management plays in reducing insider threatÔÇÖs in the federal Government", en *Information Security Journal: A Global Perspective*, vol. 31, n. 3, pp. 338–345, 2022.

[12] Alqahtani, Mohammed, Braun, Robin: "Examining the impact of technical controls, accountability and monitoring towards cyber security compliance in e-Government organisations", en *Research Square*, 2021.

[13] David G. Rosado, Julio Moreno, Luis E. Sánchez, Antonio Santos-Olmo, Manuel A. Serrano, Eduardo Fernández-Medina: "MARISMA-BiDa pattern: Integrated risk analysis for big data", en *Computers & Security*, vol. 102, pp. 102155, 2021.

[14] David G. Rosado, Antonio Santos-Olmo, Luis Enrique Sánchez, Manuel A. Serrano, Carlos Blanco, Haralambos Mouratidis, Eduardo Fernández-Medina: "Managing cybersecurity risks of cyber-physical systems: The MARISMA-CPS pattern", en *Computers in Industry*, vol. 142, pp. 103715, 2022.

[15] David G. Rosado, Luis E. Sánchez, Ángel Jesús Varela-Vaca, Antonio Santos-Olmo, María Teresa Gómez-López, Rafael M. Gasca, Eduardo Fernández-Medina: .Enabling security risk assessment and management for business process models", en *Journal of Information Security and Applications*, vol. 84, pp. 103829, 2024.

[16] Robles, Gregorio, Gamalielsson, Jonas, Lundell, Björn: "Setting up government 3.0 solutions based on open source software: the case of X-road", en *Electronic Government: 18th IFIP WG 8.5 International Conference, EGOV 2019, San Benedetto Del Tronto, Italy, September 2–4, 2019, Proceedings 18*, pp. 69–81, 2019.

[17] Odebade, Adejoke T, Benkhelifa, Elhadj: .A comparative study of national cyber security strategies of ten nations", en *arXiv preprint arXiv:2303.13938*, 2023.

[18] Srinivas Jangirala, Das Ashok Kumar, Kumar Neeraj: "Government regulations in cyber security: Framework, standards and recommendations", en *Future Gener. Comput. Syst.*, vol. 92, pp. 178–188, 2019.

[19] Carter, Lemuria, Weerakkody, Vishanth, Phillips, Brandis, Dwivedi, Yogesh K: Çitizen adoption of e-Government services: Exploring citizen perceptions of online services in the United States and United Kingdom", en *Inf. Syst. Manag.*, vol. 33, n. 2, pp. 124–140, 2016.

[20] Thompson, Nik, Mullins, Antony, Chongsutakawewong, Thanavit: "Does high e-Government adoption assure stronger security? Results from a cross-country analysis of Australia and Thailand", en *Gov. Inf. Q.*, vol. 37, n. 1, pp. 101408, 2020.

[21] Sk Tahsin Hossain, Tan Yigitcanlar, Kien Nguyen, Yue Xu: "Local Government Cybersecurity Landscape: A Systematic Review and Conceptual Framework", en *Revista/Journal*, 2023.

[22] Lino, Graduate S, Christine: Çybersecurity in the federal government: Failing to maintain a secure cyber infrastructure", en *Bull. Assoc. Inf. Sci. Technol.*, vol. 41, n. 1, pp. 24–28, 2014.

[23] Saputro Rois, Pappel Ingrid, Vainsalu Heiko, Lips Silvia, Draheim Dirk: "Prerequisites for the Adoption of the X - Road Interoperability and Data Exchange Framework: A Comparative Study.", en *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG), eDemocracy; eGovernment (ICEDEG), 2020 Seventh International Conference on*, pp. 216 - 222, 2020.

[24] Savaş, Serkan, Karataş, Süleyman: Çyber governance studies in ensuring cybersecurity: an overview of cybersecurity governance.", en *International Cybersecurity Law Review*, vol. 3, pp. 7 - 34, 2022.

[25] P. Kivimäki, "What Is a Certificate Profile?"NIIS. https://nordic-institute.atlassian.net/wiki/spaces/XRDKB/pages /16449539/What+Is+a+Certificate+Profile (accessed Nov. 5, 2024).

[26] Alrubaiq, Abdullah, Alharbi, Talal: "Developing a Cybersecurity Framework for e-Government Project in the Kingdom of Saudi Arabia", en *Journal of Cybersecurity and Privacy*, vol. 1, n. 2, pp. 302–318, 2021.

[27] Hamza Ahmad Qureshi, Yaamina Salman, Sidra Irfan, Nasira Jabeen: "A SYSTEMATIC REVIEW OF e-Government EVALUATION", en *Pakistan Economic and Social Review*, vol. 55, n. 2, pp. 355–390, 2017.

[28] Artiles, Nestor Ganuza: "State of Cybersecurity in the International Context and within NATO", en *Strategy Books*, n. 149, pp. 165–214, 2011.

[29] Abbate Janet: Ïnternet: Evolution and Challenges", en *OpenMind BBVA (docu mento en Internet http://www. bbvaopenmind. com/autor/janetabbatel)*, 2008.

[30] Bekerman, Uriel, Cresta, Guido Damián: "Digital Societies: e-Estonia (Digital Societies: e-Estonia)", en *DPI Journal, Law and Technology Supplement, Issue No.*, pp. 55–05, 2020.

[31] Zaman Noor, Shah, Imdad Ali, Rajper Samina: Çybersecurity measures for e-Government frameworks", en *Revista/Conferencia*, 2022.

Diseño y generación de un conjunto de datos para entrenamiento de modelos de prevención y detección de amenazas internas: SPEDIA dataset

David Alvarez[1], Luis Perez[1], Alberto Mateo[1], Xavier Larriva-Novol[1]
Manuel Alvarez-Campana[1], Victor A. Villagra[1]
[1]ETSI Telecomunicación, Universidad Politécnica de Madrid, España
{david.alvarez.muniz, luis.pmiguel, xavier.larriva.novo, manuel.alvarez-campana, victor.villagra}@upm.es
alberto.mateo@alumnos.upm.es

***Resumen*—Los modelos y sistemas de detección de ataques de tipo *insider* requieren de datos reales y que reflejen una variedad de comportamientos de usuario. Por los riesgos reputacionales, las empresas no publican información relativa a este tipo de incidentes, al tiempo que los datasets sintéticos disponibles públicamente suelen estar desactualizados y son difíciles de replicar.**

En este trabajo se presenta una propuesta para la generación de un conjunto de datos integrando eventos simulados en base a roles de usuario y actividades de usuarios reales participantes en un ciberejercicio de simulación de un ataque. Para ello, se define un escenario cerrado y replicable; que incluye la configuración de los modelos de simulación de comportamiento y las herramientas necesarias para monitorizar y recopilar la información necesaria para detectar el comportamiento anómalo.

***Index Terms*—Dataset, Amenazas internas, Ciberseguridad, Ejercicios de ataque**

Tipo de contribución: *Investigación en desarrollo (límite 8 páginas)*

I. Introducción

La detección de amenazas internas (Insider Threats) representa hoy en día uno de los mayores desafíos en el ámbito de la ciberseguridad. Más del 80 % de las empresas afirman sufrirlas, y por encima de la mitad de ellas las consideran más difíciles de prevenir que las amenazas externas [1]. Las amenazas relacionadas con personal interno poseen un mayor nivel de complejidad de detección. Dentro de las organizaciones, el personal interno posee legitimidad y acceso privilegiado a los sistemas y activos propios de la organización; por lo que puede ocultar su comportamiento malicioso como un acceso legítimo [2][3].

Resulta fundamental disponer de sistemas adecuados para detectar y mitigar estas amenazas. En este sentido, se considera muy atractivo el empleo de técnicas de Inteligencia Artificial. Para ello, sin embargo, es imprescindible disponer de conjuntos de datos de calidad [4][5]. El desarrollo de estos modelos debe considerar las distintas facetas del comportamiento de los usuarios dentro de una organización. Los datos requeridos no solo deben ajustarse a casos reales de ataques realizados de forma interna; también deben reflejar una diversidad de comportamientos de perfiles de usuarios dentro de una organización.

Las herramientas de detección desarrolladas en los últimos años incorporan técnicas basadas en reglas sobre patrones maliciosos previamente definidos, y técnicas basadas en el análisis de comportamiento (User Behavior Analytics) combinando modelos de aprendizaje automático y de detección de anomalías [6].

Para el desarrollo de estas herramientas se requieren registros de actividad (logs) de sistemas, tráfico de red y metadatos de uso. En la actualidad, se usan tanto datasets sintéticos, que reproducen escenarios hipotéticos o basados en ataques conocidos [7], como de fuentes reales, construidos mediante ejercicios donde usuarios "enmascarados" tienen un objetivo malicioso en un sistema [8][9].

Actualmente es difícil conseguir muestras reales de registros de actividad sin comprometer información sensible o la identidad de los empleados. Esto se debe al carácter interno de estos ataques y los riesgos económicos asociados a la reputación en ciberseguridad de una organización. Las organizaciones comúnmente no denuncian e informan sobre los ataques internos sufridos [10] y, mucho menos, publican datos de seguridad reales relativos a un incidente.

Los datasets actualmente utilizados en ámbitos científicos y considerados *benchmark datasets* sufren un desbalanceo de datos entre ejemplos de comportamiento malicioso y comportamiento legítimo, no son homogéneos en el formato de los registros de distintas fuentes y están generalmente desactualizados debido al año en que se generaron dichos ataques [6].

En este trabajo se presenta la propuesta para la generación de conjuntos de datos de amenazas internas de ciberseguridad mediante la integración de tres fuentes de datos: 1) la simulación de actividad basada en roles; 2) la recolección de actividades de usuarios reales dentro de un ciberejercicio cerrado; y 3) la incorporación de datos de actividad de datasets del estado del arte. Para ello, se ha construido un escenario cerrado y replicable, donde se monitoriza el comportamiento cotidiano de usuarios simulados y reales durante la duración del ciberejercicio.

II. Estado del arte

El desarrollo de herramientas de detección de ataques de tipo *insider* cuenta con una problemática única a resolver. No es suficiente con basar la detección en el reconocimiento de comportamiento dañino sobre un sistema, alertando sobre ejecuciones de malware y explotaciones de vulnerabilidades [2]. Se debe reconocer también el comportamiento común, individualmente indistinguible de una jornada de trabajo, que

se realiza con un objetivo malicioso. Para esto es común la aplicación de modelos de anomalías sobre representaciones básicas del comportamiento de usuario. Estos podrían suponer una ventaja para integrar la evaluación de posibles *insiders* en diferentes herramientas y sistemas de ciberseguridad.

Dado el creciente interés e importancia de identificar y mitigar los riesgos asociados con las amenazas internas en el ámbito de la ciberseguridad [11], se ha llevado a cabo una revisión exhaustiva de los estudios más recientes para determinar cuál es el dataset más utilizado por la comunidad investigadora. Este análisis no solo busca identificar el dataset de referencia, sino también comprender el tipo de información que recoge, así como las áreas que podrían no estar siendo cubiertas. Además, se pretende evaluar las principales limitaciones de estos datasets a fin de identificar sus fortalezas y debilidades en la detección de insiders.

Se ha llevado a cabo un análisis de distintos trabajos que se enfocan en el estudio de los modelos y aplicaciones más recientes en la detección de insiders. Dentro de estos se han contemplado tres trabajos [5][6][12], desarrollados como revisiones del estado del arte. Estos trabajos exploran diversos enfoques para la detección de amenazas internas, centrándose en técnicas de machine learning y deep learning, así como en el análisis de los datasets empleados para entrenar y evaluar los modelos.

Los tres estudios considerados coinciden en la prevalencia del empleo de los dataset CERT [7], en sus distintas versiones, frente a otros conjuntos de datos, tal como se representa en la **Figura 1**. Así, en [5], se indica que el dataset CERT se utilizaba en el 52,63 % de los trabajos analizados. Los otros dos estudios [6] [12], más recientes, mostraban un uso aún mayor, del el 75 % y el 74,19 %, respectivamente.

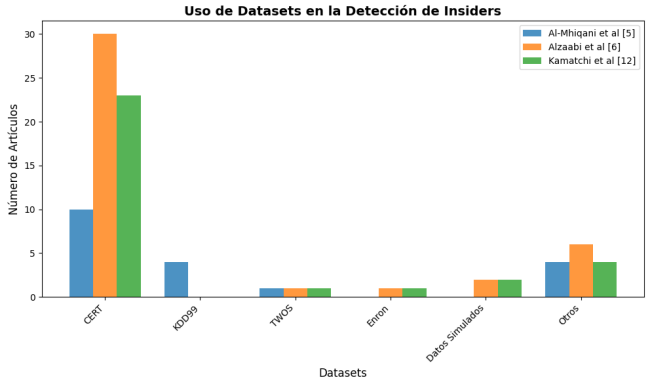


Figura 1. Uso de datasets en la detección de insiders

El dataset CERT, en su version 4.2, es ampliamente utilizado en la detección de amenazas internas debido a que ofrece un conjunto de datos bien estructurado con atributos claramente definidos. Estas características lo hacen ideal para la validación y prueba de modelos de detección de amenazas internas. La naturaleza accesible de este dataset ha facilitado su adopción en numerosos estudios e investigaciones. Sin embargo, uno de los problemas más importantes al trabajar con este dataset es la falta de información detallada sobre el escenario específico en el que los datos fueron generados. Aunque los atributos que se recogen están claramente especificados, como el inicio y final de sesiones, conexiones USB, contenido web (URLs y contenido de las páginas accedidas), correos electrónicos y archivos, no se ofrece ninguna explicación sobre cómo se llevó a cabo la generación.

Esto representa un desafío significativo para quienes deseen replicar el dataset, ya que, aunque los atributos son comprensibles, es necesario definir un escenario para la recolección de datos similares. La replicación de los datos no es tan sencilla como simplemente definir un escenario y registrar la información de las actividades. De hecho, replicar todo el proceso de recolección de datos sería sumamente complejo debido a la infraestructura y las herramientas necesarias para monitorizar los diferentes tipos de actividades.

Por ejemplo, la recolección de información de navegación web, como las URLs y el contenido de las páginas accedidas, presenta varios desafíos técnicos. Si se utiliza HTTPS para cifrar las comunicaciones, los datos estarán protegidos por TLS, lo que dificulta su captura. Para acceder al contenido real de una página web, sería necesario desarrollar un scraper específico que pueda manejar las complejidades de TLS y HTTPS, lo cual no es trivial. Otra complejidad adicional es la interpretación de páginas web que incluyen scripts que generan contenido dinámico.

Otro aspecto igualmente desafiante es la monitorización de las conexiones USB. Si bien puede parecer relativamente sencillo registrar cuándo se conecta un dispositivo USB, monitorizar en tiempo real cualquier cambio o actividad dentro de esos dispositivos es mucho más complejo. La necesidad de registrar cambios en los archivos, accesos y modificaciones dentro de los dispositivos USB requiere un sistema robusto que pueda manejar la detección y notificación de estos eventos. Además, la heterogeneidad con la que se gestionan los sistemas operativos y exponen los eventos relacionados con los dispositivos USB, hace aún más difícil la tarea de replicar este proceso en diversos entornos.

Un problema adicional importante del dataset CERT es el desbalanceo extremo entre las clases, lo que puede llevar a resultados sesgados e imprecisos en los modelos de detección [13]. Esta desproporción no se aborda adecuadamente en la mayoría de los enfoques actuales, lo que dificulta la eficacia de los modelos entrenados con este dataset. Las clases desbalanceadas pueden provocar que los modelos favorezcan las clases más representadas, reduciendo así su capacidad para detectar con precisión las amenazas internas.

Cabe destacar que el dataset CERT fue desarrollado por el Software Engineering Institute (SEI) de la Carnegie Mellon University, publicándose por primera vez en 2014 como parte de un esfuerzo por proporcionar datos para el estudio de amenazas internas. A pesar de haberse convertido en el dataset referente en el campo de la detección de amenazas internas, no ha sido actualizado desde 2020, fecha en la que se publicó su última versión [14]. Esto implica que muchos de los comportamientos y escenarios que refleja podrían no alinearse con las prácticas actuales, limitando la capacidad de los modelos para adaptarse a amenazas emergentes. Así, aunque sigue siendo una referencia clave, su antigüedad y falta de actualización subrayan la necesidad de datasets más recientes que capturen dinámicas contemporáneas en entornos

reales de trabajo.

Otro dataset comúnmente utilizado en los últimos años es el dataset TWOS (Trusted Wireless Observation System) [8], el cual se distingue por no ser completamente simulado, ya que sus datos provienen de un entorno controlado y definido específicamente para la recolección de información en escenarios de amenazas internas. A diferencia del dataset CERT, el TWOS recoge actividades reales realizadas en un entorno de red supervisado, lo que le otorga un mayor nivel de autenticidad para representar el comportamiento de usuarios potencialmente maliciosos. Entre los tipos de datos recopilados se encuentran la actividad del ratón, pulsaciones de teclado, actividades en el sistema anfitrión (como llamadas al sistema o interacción con el sistema de archivos), tráfico de red, actividad de correo electrónico, registros de inicio y cierre de sesión, y comandos UNIX. Esta diversidad de fuentes de información ofrece una visión más completa del comportamiento del usuario y permite detectar patrones que podrían pasar desapercibidos en datasets más simplificados.

A pesar de sus ventajas, el TWOS no ha sido tan adoptado en los estudios recientes. Esto podría deberse a que no cuenta con la misma visibilidad que el dataset CERT, y a que la comunidad investigadora ha tendido a preferir datasets con estructuras más estandarizadas y accesibles. Asimismo, el TWOS presenta ciertos desafíos de integración, ya que la diversidad de fuentes de datos que ofrece requiere un preprocesamiento más complejo para unificar la información antes de entrenar los modelos.

III. Metodología

La metodología que se propone para la generación del dataset se basa en la definición de un entorno en el que se combinan eventos simulados y usuarios reales participando en un un ejercicio de ciberseguridad. El formato del dataset a generar sigue, pero expande, el utilizado por el del CERT. Para ello, se amplían los registros mediante la integración de tres perfiles de actividad: 1) actividad de usuarios reales derivada de la ejecución de técnicas, tácticas y procedimientos llevados a cabo por participantes en un ciberejercicio enfocado en ejecutar un ataque interno; 2) eventos realizados por usuarios simulados a partir de la generación pseudoaleatoria de actividad normal (no maliciosa), basada en un modelo de comportamiento; y 3) el comportamiento de usuarios sintéticos replicados a partir del dataset CERT [7].

Para la construcción, configuración y ejecución del entorno del ciberejercicio y la simulación de los que se recogen los datos se han segudi los siguientes pasos:

1. **Definición del escenario**. El escenario consiste en tres máquinas virtuales en las que se ejecutan tareas tanto por los usuarios simulados como por los participantes en el ciberejercicio. En la Sección IV se describe la configuración y servicios disponibles en las máquinas, así como la estructura de red del ejercicio. También se detallan los posibles roles que pueden tomar los usuarios simulados en cada máquina.
2. **Creación del modelo de comportamiento**. Se han definido los usuarios simulados, su rol en el equipo, las máquinas y servicios a los que pueden acceder, así como sus patrones de trabajo. Para cada usuario se han propuesto patrones de comportamiento como bloques de comandos que pueden ser realizados mediante una conexión *ssh*, y que permiten el uso, configuración y mantenimiento de los servicios activos en las máquinas. En la Sección V-A se detalla la implementación seguida para crear esta simulación.
3. **Integración de datos CERT**. Entre los comandos realizados al seguir el modelo de comportamiento se ha querido incluir registros replicando la actividad del dataset CERT. En concreto, replicando registros de acceso a *urls*, envió de correos, uso de dispositivos extraibles y modificación de ficheros. Para esto, en el modelo de comportamiento se ha implementado la selección por *sampling* sobre los registros no maliciosos incluidos en este dataset; envíando una copia del registro seleccionado al sistema central de almacenamiento.
4. **Ejecución del ciberejercicio**. En copias locales pero monitorizadas del sistema definido en el escenario, se solicita a los participantes la ejecución de un escenario de ataque de tipo *insider* durante un periodo máximo de 30 días.
5. **Simulación de actividad no maliciosa**. Sobre una copia central del sistema definido en el escenario durante la duración total del ciberejercicio se mantiene en constante funcionamiento una simulación. Esta simulación ejecuta los comandos marcados por el modelo de comportamiento haciéndose pasar por cada usuario simulado.
6. **Limpieza y procesado del dataset**. Finalmente, se construye un dataset como resultado del ejercicio. Se propone tanto un formato dataset en bruto, constituido de las alertas generadas por el sistema de gestión de eventos monitorizando las máquinas incluidas en el ejercicio. Como una versión simplificada que sigue el formato definido en el CERT. Agregando la actividad de usuario proveniente de la creación o borrado de un archivo, ejecución de comandos y uso de servicios durante una sesión.

IV. Escenario propuesto

Se propone un escenario base a la imagen de un ataque de tipo insider realizado por un ex-empleado de perfil técnico. El escenario recrea un entorno que permita un ataque llevado a cabo por dicho ex-empleado que busca acceder a la red interna privada de la organización con el fin de obtener información confidencial.

Este ex-emplead, desde su ordenador personal, sigue teniendo acceso a algunos de los recursos comunes de la empresa objetivo. Al no realizar un proceso correcto de terminación, eliminando accesos y privilegios al romper contacto con clientes y usuarios, una empresa se expone a que este acceso privilegiado sea usado como acceso inicial (Initial Access (TA0001)) [15].

En base a este acceso inicial, el ex-emplead va a realizar un movimiento lateral a un servidor interno. Un activo de mayor riesgo, solo accesible mediante una red interna; cuya pérdida de confidencialidad supone un impacto grave para la organización.

El entorno experimental está entonces compuesto por tres máquinas principales, tal y como se puede observar en la **Figura 2**, cada una de estas con un rol específico dentro de



Figura 2. Escenario propuesto

la arquitectura de la red. Además, sobre cada una se declara un conjunto cerrado de roles de usuario que operan sobre estas máquinas. Estos roles serán usados para definir cómo interactúan con el escenario los modelos de simulación; y para definir las cuentas que puede usar un atacante para realizar escalada de privilegios y movimientos laterales.

1. **Máquina atacante:** Representa el ordenador personal del ex-empleado malintencionado. Está conectada a una red externa denominada *extranet*, desde donde intentará comprometer a los otros sistemas. Esta máquina cuenta con herramientas de explotación y análisis de vulnerabilidades, incluyendo escáneres de puertos, sniffers de tráfico y software para ataques de fuerza bruta o ingeniería social. Pues el ex-emplead cuenta con acceso administrativo sobre la máquina. Por lo tanto contamos con único usuario:
   - **Ex-empleado atacante.** Con un acceso completo a los recursos de su máquina.
2. **Host accesible:** Simula un periferico expuesto a la *extranet* que sera utilizado como punto de entrada a la red interna. En este caso, se configura como una impresora de red, un dispositivo que a menudo tiene medidas de seguridad mínimas y es susceptible a ataques de configuración incorrecta o firmware desactualizado. Además, este host cuenta con un acceso limitado a la *intranet*, la red interna de la organización desde la que es accesible el activo objetivo. Esto permite al atacante intentar moverse lateralmente dentro de la red.
   Esta máquina tiene tres puertos abiertos que cumplen funciones comunes en un entorno empresarial:
   - **Puerto 21 (FTP):** Permite la transferencia de archivos a la impresora, lo que facilita la carga de documentos para su posterior impresión o almacenamiento en el dispositivo.
   - **Puerto 22 (SSH):** Habilita la gestión remota del dispositivo, permitiendo a los administradores acceder de manera segura para realizar tareas de mantenimiento, configuración y actualizaciones.
   - **Puerto 80 (HTTP):** Aloja un servidor web en el que corre una aplicación denominada *PHP Charts*, utilizada para visualizar estadísticas relacionadas con el uso del dispositivo y la gestión de trabajos de impresión.

   Sobre este sistema definimos dos roles de usuario:
   - **Administrador de sistemas:** Encargado de dar soporte y mantenimiento al dispositivo. Posee credenciales con mayores privilegios, lo que le permite configurar opciones avanzadas (direcciones de red, actualizaciones de firmware, etc.) y gestionar el acceso de otros usuarios.
   - **Usuario del dispositivo:** Empleado de la organización que utiliza el dispositivo para tareas cotidianas de impresión o escaneo. Tiene permisos básicos limitados a las funciones propias de impresión.
3. **Servidor interno:** Este servidor se encuentra dentro de la *intranet* de la empresa y almacena información sensible, como registros de empleados, documentos corporativos y bases de datos confidenciales. La protección de dicho servidor se basa en la autenticación mediante credenciales y en la segmentación de la red. Sin embargo, en él se ejecutan servicios que podrían ser explotados por un atacante con acceso a credenciales privilegiadas o a través de vulnerabilidades presentes en el software.
   Este servidor tiene dos puertos abiertos fundamentales en un entorno empresarial:
   - **Puerto 22 (SSH):** Permite la administración remota del servidor, proporcionando un acceso seguro a los administradores del sistema para llevar a cabo configuraciones, actualizaciones y mantenimiento.
   - **Puerto 80 (HTTP):** Alberga un servidor web en el que se ejecuta el servicio de correo interno de la organización mediante *Apache2*. Este servicio es clave para la comunicación interna y la gestión de correos electrónicos dentro de la empresa

   Con respecto a los usuarios, se consideran con tres roles:
   - **Administrador del servicio:** Encargado de supervisar el funcionamiento y uso de estos y otros servicios activos en la máquina, ofreciendo el soporte necesario cuando se presentan fallos.
   - **Desarrolladores web:** Responsables de las modificaciones básicas de la funcionalidad ofrecida en el servidor web. Cuentan con acceso para editar la configuración del servidor web y para modificar la implementación web, pero no poseen permisos sobre la configuración y mantenimiento de los servicios del sistema que sustentan el funcionamiento de las aplicaciones web.
   - **Gestor de cuentas:** Responsable de la configuración de cuentas, perfiles y accesos al servicio web. Esto supone un permiso en la modificación de las bases de datos que guardan y controlan la información privilegiada a la que pretende acceder el usuario atacante.

Para garantizar la coherencia del escenario, la infraestructura de red se ha configurado con parámetros que simulan un entorno empresarial realista. Se han definido restricciones de conectividad entre las distintas máquinas, asegurando que el atacante sólo pueda interactuar inicialmente con el host accesible y deba encontrar métodos para moverse lateralmente dentro de la red.

El diseño del entorno también contempla diferentes niveles de privilegios en las cuentas de usuario dentro de las máquinas comprometidas. Esto permite que el atacante deba emplear técnicas como el análisis de tráfico en la red, la explotación de vulnerabilidades en servicios expuestos y el uso de credenciales filtradas para acceder a sistemas con mayores permisos.

V. Explotación del escenario

Sobre el escenario definido se aplican dos modos de explotación para la creación de datos, asociados a los perfiles de actividad mencionados en la Sección III. Por un lado, la simulación de comportamiento no malicioso, incluyendo actividad simulada y replicada del dataset CERT; y por otro, la ejecución del ciberejercicio.

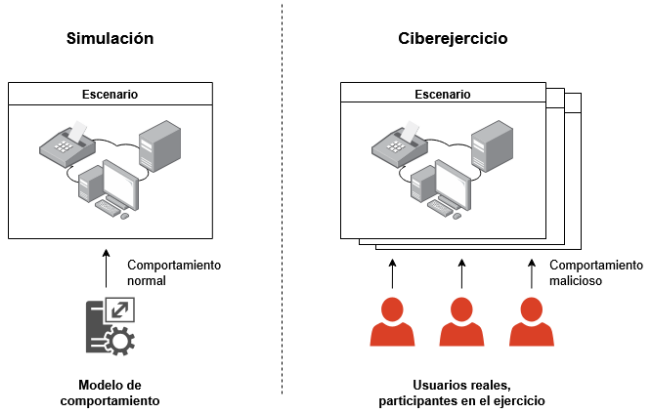


Figura 3. Modelos de explotación.

V-A. *Modelo de comportamiento*

Con el propósito de ocultar el comportamiento malicioso en el dataset propuesto, se simularán jornadas de trabajo realistas en las máquinas que conforman el escenario. Es esencial que el tipo de comportamiento y el tráfico reflejados se asemejen a los utilizados por los atacantes. Para ello, se reproducirá el uso, mantenimiento y configuración de los servicios clave de dichas máquinas, los mismos que podrían ser explotados por los atacantes.

Para diseñar el flujo de trabajo de los usuarios, se parte de un modelo definido a partir de dos perfiles: el rol de trabajo y el patrón de comportamiento. El **rol de trabajo** determina las tareas que un usuario puede ejecutar en las máquinas a las que tiene acceso. Por ejemplo, un administrador de sistemas podría realizar, entre otras acciones:

- Configurar, actualizaciones y gestionar servicios críticos.
- Consultar y analizar los *logs de servicios*.
- Simular el uso de los servicios para comprobar su funcionamiento.

Las interacciones asociadas a cada rol se plasman en "bloques de trabajo" que un usuario ejecuta a lo largo de su jornada. Cada bloque se implementa como un conjunto de instrucciones o comandos que el usuario introduce, a través de conexiones *SSH* gestionadas mediante la librería *paramiko*, para operar sobre los servicios de la máquina.

Por ejemplo un bloque asociado a un cambio sobre la configuración del servicio *vsftpd*, incluyendo la exclusión de una ip se define como el siguiente conjunto de comandos:

```
$ sudo sed -i '/tcp_wrappers=/d'
/etc/vsftpd.conf
$ sudo sh -c "echo 'tcp_wrappers=YES' >>
/etc/vsftpd.conf"
$ sudo sh -c "echo 'vsftpd: 192.168.1.' >>
/etc/hosts.allow"
$ sudo -S systemctl --no-pager restart vsftpd
$ systemctl --no-pager status vsftpd
```

El **patrón de comportamiento** de cada usuario describe la forma en que se llevan a cabo dichas tareas a lo largo de la jornada. Esta se compone de la ejecución pseudoaleatoria (respetando los rangos indicados en el patrón de comportamiento) de sesiones *SSH*, durante las cuales se realizan, de manera consecutiva, los diferentes bloques de trabajo. Para cada usuario, se definen los siguientes parámetros:

- Horario laboral durante el cual se pueden iniciar sesiones *ssh*.
- Rangos temporales para los descansos entre cada sesión *ssh*.
- Cantidad de bloques de trabajo que se agrupan en una sesión *ssh*.
- Retardo entre bloques dentro de una misma sesión.
- Porcentaje de registros procedentes del cert replicados durante una sesión ssh.

Estos "bloques de trabajo" surgen al combinar las responsabilidades asociadas a cada rol (por ejemplo, administrador de sistemas) con los servicios disponibles en la máquina (por ejemplo, *vsftpd*, *ssh*, *apache2*).

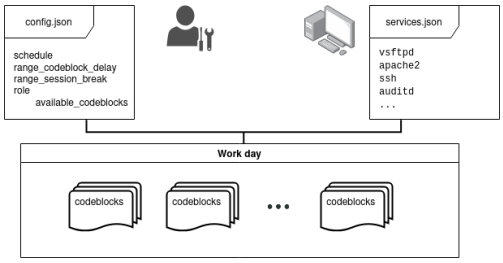


Figura 4. Modelo de comportamiento.

Con base en el modelo de comportamiento definido, se mantiene en ejecución constante una instancia central del escenario. Allí, los usuarios simulados llevarán a cabo jornadas laborales sintéticas a lo largo de los 30 días destinados a la realización del ciberejercicio descrito en la siguiente sección.

V-B. *Ejecución del ciberejercicio*

El ciberejercicio se desarrolla en una actividad práctica donde usuarios reales con un conocimiento técnico en telecomunicaciones y ciberseguridad disponen de un periodo de 30 días para ejecutar un ataque. Para ello, cada usuario participante operará sobre una copia local del escenario planteado. Cada copia local se monitoriza mediante el uso de agentes de un sistema central de Gestión de Eventos e Información de Seguridad (SIEM); un sistema propio en desarrollo para el proyecto SPEDIA [16], bajo el que también se ha realizado este trabajo.

El ataque se realiza en varias fases que reflejan las tácticas y técnicas descritas en la matriz MITRE ATT&CK [17], imitando el comportamiento de un atacante real con conocimiento previo de la infraestructura de la empresa. En primer lugar, el acceso inicial se corresponde con *Initial Access (TA0001)* mediante el escaneo de servicios expuestos y la explotación de vulnerabilidades *(Exploit Public-Facing Application - T1190)*. A continuación, la ejecución de *exploits*

y el establecimiento de persistencia se asocian con *Execution (TA0002)* y *Persistence (TA0003)*, respectivamente. Seguidamente, la escalada de privilegios se clasifica en *Privilege Escalation (TA0004)*, mientras que los movimientos laterales se vinculan con *Lateral Movement (TA0008)* por medio del acceso a otros sistemas utilizando *Remote Services: SSH (T1021.004)*. Por último, el compromiso del servidor y la extracción de información corresponden a *Credential Access (TA0006)* y *Exfiltration (TA0010)*, completando de este modo un ciclo de ataque coherente con las TTPs de MITRE.

1. **Acceso inicial:** El atacante busca vulnerabilidades en el host accesible para establecer un primer punto de acceso. Para ello, realiza un análisis perimetral de la red utilizando herramientas como Nmap y Metasploit, permitiendo mapear la infraestructura y detectar servicios expuestos que puedan ser explotados. Además, analiza las respuestas del sistema a solicitudes específicas para obtener información adicional sobre las versiones de software en ejecución y posibles configuraciones inseguras.
2. **Explotación de vulnerabilidades:** Una vez identificado un servicio vulnerable en el host accesible, el atacante encuentra un exploit adecuado dentro de Metasploit, lo configura y lo ejecuta para obtener acceso al sistema comprometido. A partir de este punto, se procede a la exploración de archivos y credenciales almacenadas en el sistema. En este proceso, el atacante examina registros del sistema, configuraciones y posibles archivos con información sensible que puedan facilitar el acceso a otros sistemas dentro de la red.
3. **Escalada de privilegios:** Con acceso al host accesible, el atacante intenta obtener credenciales de otros usuarios con mayores permisos. Se conecta al servicio FTP de la máquina, donde encuentra archivos protegidos con contraseñas. Mediante herramientas como John the Ripper, descifra las credenciales y accede a nuevos recursos dentro de la red interna.
4. **Movimientos laterales:** El atacante emplea técnicas de pivotaje y túneles de red para moverse lateralmente y acceder a la intranet de la organización. Una vez dentro, compromete el servidor interno y busca información sensible almacenada en él. Para ello, puede utilizar herramientas de enumeración de sistemas y protocolos de acceso remoto como SSH. Además, analiza posibles configuraciones erróneas que permitan el acceso a otros sistemas críticos.
5. **Exfiltración de datos y persistencia:** Finalmente, el atacante extrae archivos específicos a través de conexiones de red, dispositivos USB o correos electrónicos. Además de obtener documentos confidenciales, en esta fase el atacante puede establecer mecanismos para mantener su acceso en el tiempo, como la creación de usuarios ocultos, la modificación de configuraciones del sistema y la instalación de puertas traseras que le permitan reconectarse en caso de que su acceso inicial sea revocado.

VI. Recolección de los datos

Como se mencionó en la anterior sección, el escenario está siendo monitorizado mediante agentes de SPEDIA tanto para la simulación como para el ciberejercicio. Con el fin de recolectar un conjunto de datos más completo y representativo, se realizaron modificaciones en estos agentes para ampliar su capacidad de detección y registrar actividades que reflejan distintas fases de un ataque interno. Estas mejoras permitieron capturar información detallada sobre los siguientes vectores de ataque y comportamientos en el sistema:

- **Conexiones FTP:** Se implementó la monitorización de sesiones FTP, registrando autenticaciones, transferencias de archivos y comandos ejecutados. Esto permite la identificación de posibles intentos de acceso no autorizado o exfiltración de información.
- **Conexiones SSH:** Se ha configurado el registro de las conexiones, fallidas o completas, realizadas mediante *SSH*.
- **Ejecución de comandos:** Se incorporo la auditoría de todos los comandos ejecutados por usuarios, y no por el sistema, en las tres máquinas mediante el uso de *auditd*. Además, se ha definido un listado de comandos de especial interés ordenados en tres niveles de alerta.
- **Integridad de ficheros:** Se activo la monitorización de cambios, creaciones y eliminaciones de ficheros por parte de los usuarios. Esto permite identificar el usuario a cargo de cambio y el tamaño antes y despues de cada modificación.
- **Navegación web:** Se configuró la detección de accesos a páginas web y el análisis de tráfico HTTP/HTTPS, lo que facilita la identificación de posibles comunicaciones con servidores externos o sitios sospechosos.
- **Uso de servicios:** Se ha configurado el registro de los usos, o intentos de uso, de cada puerto donde se desplegaron los servicios *FTP, HTTP* o *SSH*.
- **Procesos en activo:** Se han definido escaneos periódicos para identificar y registrar la ejecución de posibles procesos maliciosos en activo. En el listado de estos procesos maliciosos se ha incluido herramientas comunes de escaneo de puertos (*nmap*, *masscan*), y de creación de tuneles y shells inversos (*netcat*, *socat*).

VII. Dataset

Tras la realización del escenario propuesto previamente por los participantes del ciberejercicio durante un periodo de 30 días, se obtuvieron todos los registros que posteriormente serán utilizados para conformar los datasets.

Respecto al proceso de limpieza del conjunto de datos, se aplican dos enfoques diferenciados. El primero aborda el tratamiento de las alertas recopiladas directamente por el sistema SPEDIA, conservando inicialmente un alto nivel de detalle debido al gran volumen de información obtenida del sistema durante el ejercicio. Posteriormente, se seleccionaron aquellas variables que sintetizan de forma más efectiva la información relevante para este estudio, logrando reducir las 202 columnas originales a tan solo 25. Las columnas eliminadas estaban compuestas en su totalidad por información interna del sistema SPEDIA, tales como las reglas empleadas, la estructura interna de almacenamiento asociada a los índices o identificadores únicos de los eventos. Además, se realizó una revisión exhaustiva para garantizar la consistencia y corrección de los datos recopilados, tratando adecuadamente

los valores nulos y verificando que cada campo contenga la información correspondiente.

La **Tabla I** muestra las 25 columnas mencionadas anteriormente, organizadas según la naturaleza de la información recopilada. Las primeras cuatro filas de la tabla están destinadas a la identificación del usuario: inicialmente aparece *"Agent_name"*, seguido por *"Dstuser"* y *"Syscheck_name"*, campos responsables de identificar al usuario que realiza cualquier actividad registrada durante el ciberejercicio. Por otro lado, se incluye *"User"*, que especifica el nombre del usuario asociado a los registros de eventos del CERT.

Desde la columna 5 hasta la columna 9, se recopila información relacionada con el sistema, incluyendo el tiempo exacto en que se registró el evento *"Timestamp"*, el nombre del decodificador utilizado *"Decoder_name"*, la descripción detallada del evento "Description", la dirección IP de origen *"Srcip"* y el registro completo del evento *"Full_log"*.

Las dos siguientes columnas indican el tipo de actividad registrada según el origen del evento; *"Activity_CERT"* para los eventos provenientes del CERT y *"Activity_SIEM"* para los registros generados por las acciones de los usuarios. Las columnas restantes proporcionan detalles específicos asociados a cada tipo de actividad registrada. Así, para eventos del tipo *"Url"*, se incluyen los campos *"Content"* y *"Url"*; para eventos relacionados con *"Email"*, se recopilan campos como *"Content"*, *"To"*, *"Cc"*, *"Bcc"*, *"From"*, *"Attachments"* y *"Size"*. Por otra parte, los eventos relacionados con *"File"* incluyen los campos *"Content"*, *"Filename"*, *"Path"*, *"Size"*, *"Size_after"* y *"Size_before"*. Por último, el nuevo tipo de actividad *"Command"* se completa por el campo *"Level"* y *"Command"*.

El segundo enfoque sigue una estructura basada en el modelo del conjunto de datos CERT, pero incorpora variables adicionales obtenidas durante la ejecución del ciberejercicio. Dichas variables adicionales contemplan eventos como la creación o eliminación de archivos, la ejecución específica de comandos y el uso de servicios por parte de los usuarios durante la jornada laboral. En consecuencia, fue necesario integrar la información simulada procedente del dataset CERT con los datos reales recopilados durante el ejercicio, logrando unificar ambos conjuntos bajo el formato original del dataset CERT e incorporando simultáneamente los nuevos campos obtenidos de *"Action"*, *"Command"* y *"Level"*.

Este resultado puede observarse en la **Tabla II**, en la cual se utiliza la estructura definida por los campos de la **Tabla I** para alcanzar una representación homogénea que no diferencia entre información simulada e información recopilada de las acciones reales realizadas por los participantes del ciberejercicio. Esto queda ejemplificado con el campo *"User"*, encargado de identificar al usuario que realiza cualquier tipo de actividad y representado en la **Tabla II** mediante los campos *"Dstuser"*, *"Syscheck_name"* y *"User"*. Por su parte, el campo *"Hostname"* queda reflejado en la **Tabla II** mediante el campo *"Agent_name"*.

Finalmente, el campo *"Activity"* unifica ahora todos los tipos de actividad registrados durante el ejercicio, mientras que las columnas restantes recogen información adicional específica relacionada con cada tipo de actividad, siguiendo el formato descrito en la **Tabla I**.

Los conjuntos de datos disponibles, en su versión preliminar, se encuentran en el siguiente enlace: `https://drive.upm.es/s/KxpQuJvxfd2psjv`. Como parte de la propuesta presentada, el ciberejercicio continúa en desarrollo; por tanto, la información se actualizará de forma progresiva hasta presentar su primera versión oficial en Zenodo [18].

Tabla I
RAW DATASET CARACTERÍSTICAS

#	Nombre	Descripción
1	Agent_name	Nombre del agente que generó el registro .
2	Dstuser	Nombre del usuario que realiza eventos de tipo: Command, Url, LogOn, LogOff.
3	Syscheck_name	Nombre del usuario que realiza una acción de tipo File.
4	User	Nombre del usuario que generó un evento simulado.
5	Timestamp	Fecha y hora en que se registró el evento.
6	Decoder_name	Nombre del decodificador que interpretó el log.
7	Description	Descripción de los eventos recogidos por el SIEM.
8	Srcip	Dirección IP de origen desde la que se generó el evento.
9	Full_log	Registro completo del log del evento.
10	Activity_CERT	Tipo de evento simulado: Email, File, Url, Connect, Disconnect.
11	Activity_SIEM	Tipo de evento recogido por SIEM: File, Url, Command, LogOn, LogOff.
12	Content	Contenido específico asociado a la actividad Url, File o Email, como el texto interno de un correo electrónico o los datos de un archivo.
13	Url	Dirección URL asociada al evento.
14	To	Dirección de correo a la que se envía la información.
15	Cc	Destinatarios en copia visible del mensaje.
16	Bcc	Destinatarios en copia oculta del mensaje.
17	From	Remitente del email.
18	Attachments	Número de archivos adjuntos asociados al evento.
19	Size	Tamaño del contenido del archivo asociado al evento de tipo File o Email.
20	Size_before	Tamaño del contenido del archivo antes de realizar una acción.
21	Size_after	Tamaño del contenido del archivo tras realizar una acción.
22	Filename	Nombre del archivo involucrado en el evento.
23	Path	Ruta del sistema donde se localiza el archivo.
24	Command	Nombre del comando ejecutado por el usuario.
25	Level	Nivel de riesgo del comando ejecutado.

VIII. DISCUSIÓN Y LINEAS FUTURAS

Durante el desarrollo del proyecto SPEDIA [16], bajo el que se ha realizado este estudio, se ha propuesto una herramienta integrada para la detección y prevención de ataques *insider*; que incorpore alertas de distintos orígenes para evaluar el riesgo general en una organización. La falta de datos reales y la poca homogeneidad entre los datasets disponibles en el estado del arte dificultan el desarrollo y sobre todo la evaluación de estas herramientas.

Uno de los objetivos al desarrollar este conjunto de datos era poder extender los modelos de detección de anomalías entrenados sobre las variantes del dataset CERT. Estos modelos operan como un clasificador sobre las sesiones de usuarios. Cada sesión inicia con un evento de inicio de sesión "LogOn"

Tabla II
CERT DATASET CARACTERÍSTICAS

#	Nombre	Descripción
1	User	Nombre del usuario que generó el evento.
2	Hostname	Nombre del equipo desde donde se generó el evento.
3	Timestamp	Fecha y hora en que se registró el evento.
4	Activity	Tipo de evento: Email, File, Url, Command, Connect, Disconnect LogOn, LogOff.
5	Url	Enlace asociado al evento.
6	Content	Contenido específico asociado a la actividad Url, File o Email, como el texto interno de un correo electrónico o los datos de un archivo.
7	To	Dirección de correo a la que se envía la información.
8	Cc	Destinatarios en copia visible del mensaje.
9	Bcc	Destinatarios en copia oculta del mensaje.
10	From	Remitente del evento.
11	Attachments	Número de archivos adjuntos asociados al evento.
12	Size	Tamaño del contenido del archivo asociado al evento de tipo File o Email.
13	Filename	Nombre del archivo asociado al evento.
14	Action	Tipo de acción sobre archivo: Change, Add o Delete.
15	Command	Nombre del comando ejecutado por el usuario.
16	Level	Nivel de severidad o criticidad del evento.

y concluye con un cierre de sesión "LogOff". Por consiguiente, todos los eventos ocurridos dentro de este intervalo quedan asociados a dicha sesión particular del usuario. Además, añadiendo información contextual relevante, incluyendo el rol de usuario dentro de la organización, correo electrónico y departamento, entre otros atributos relacionados con su posición dentro de la empresa.

La propuesta actual amplía esta representación mediante un mayor nivel de detalle en eventos existentes o la inclusión de nuevos tipos de eventos que pueden registrarse durante una sesión. En concreto, se agrega información sobre acciones realizadas sobre archivos y comandos ejecutados en las máquinas de los usuarios, presentados en la **Tabla II**. Estos aspectos son especialmente relevantes para la detección y análisis de amenazas internas, ya que permiten categorizar los comandos ejecutados según distintos niveles de riesgo o el tipo de acción realizada sobre un archivo, facilitando así su integración en vectores de conocimiento que potencien la identificación de posibles insiders dentro de la organización.

La metodología propuesta para la generación de datos *insider* sigue sufriendo de algunos de los problemas identificados en el estado del arte. Los datos normales, no maliciosos, son sintéticos; y han sido creados siguiendo modelos de comportamiento que ejecutan actividades de muy poco riesgo. La actividad realizada por usuarios reales en una organización es más caótica y puede incluir errores o cambios en el desarrollo que se asemejan más al comportamiento del atacante.

Aún así, esta propuesta no pretende solamente generar otro conjunto de datos. El principal objetivo es marcar una metodología que podamos reutilizar para la creación de otros ejercicios similares que ofrezcan datos reales asociados a otros escenarios. Esto permite evaluar la generalidad de los modelos de detección y las herramientas de prevención contra ataques internos.

AGRADECIMIENTOS

Este trabajo está cofinanciado por el Instituto Nacional de Ciberseguridad de España (INCIBE) a través del programa de promoción de Proyectos Estratégicos de Ciberseguridad en España, enmarcado dentro del Plan de Recuperación, Transformación y Resiliencia, mediante fondos del programa Next Generation EU.

REFERENCIAS

[1] C. Insiders, "2024 insider threat report," 2024, consultado el 03 de marzo de 2025. [Online]. Available: https://gurucul.com/2024-insider-threat-report/
[2] J. Kim, M. Park, H. Kim, S. Cho, and P. Kang, "Insider threat detection based on user behavior modeling and anomaly detection algorithms," *Applied Sciences*, vol. 9, no. 19, p. 4018, 2019.
[3] S. Asha, D. Shanmugapriya, and G. Padmavathi, "Malicious insider threat detection using variation of sampling methods for anomaly detection in cloud environment," *Computers and Electrical Engineering*, vol. 105, p. 108519, 2023.
[4] T. Rashid, I. Agrafiotis, and J. R. Nurse, "A new take on detecting insider threats: exploring the use of hidden markov models," in *Proceedings of the 8th ACM CCS International workshop on managing insider security threats*, 2016, pp. 47–56.
[5] M. N. Al-Mhiqani, R. Ahmad, Z. Zainal Abidin, W. Yassin, A. Hassan, K. H. Abdulkareem, N. S. Ali, and Z. Yunos, "A review of insider threat detection: Classification, machine learning techniques, datasets, open challenges, and recommendations," *Applied Sciences*, vol. 10, no. 15, p. 5208, 2020.
[6] F. R. Alzaabi and A. Mehmood, "A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods," *IEEE Access*, vol. 12, pp. 30 907–30 927, 2024.
[7] J. Glasser and B. Lindauer, "Bridging the gap: A pragmatic approach to generating insider threat data," in *2013 IEEE Security and Privacy Workshops*. IEEE, 2013, pp. 98–104.
[8] A. Harilal, F. Toffalini, J. Castellanos, J. Guarnizo, I. Homoliak, and M. Ochoa, "Twos: A dataset of malicious insider threat behavior based on a gamified competition," in *Proceedings of the 2017 international workshop on managing insider security threats*, 2017, pp. 45–56.
[9] H. A. Kholidy and F. Baiardi, "Cidd: A cloud intrusion detection dataset for cloud computing and masquerade attacks," in *2012 Ninth International Conference on Information Technology-New Generations*. IEEE, 2012, pp. 397–402.
[10] F. Whitelaw, J. Riley, and N. Elmrabit, "A review of the insider threat, a practitioner perspective within the uk financial services," *IEEE Access*, vol. 12, pp. 34 752–34 768, 2024.
[11] G. Sandra, S. Silas, and E. B. Rajsingh, "Assessing current scenario of insider threat in the perspective of recent pandemic," in *2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*. IEEE, 2024, pp. 1–7.
[12] K. Kamatchi and E. Uma, "Insights into user behavioral-based insider threat detection: systematic review," *International Journal of Information Security*, vol. 24, no. 2, p. 88, 2025.
[13] T. Al-Shehari and R. A. Alsowail, "Random resampling algorithms for addressing the imbalanced dataset classes in insider threat detection," *International Journal of Information Security*, vol. 22, no. 3, pp. 611–629, 2023.
[14] B. Lindauer, "Insider Threat Test Dataset," 9 2020. [Online]. Available: https://kilthub.cmu.edu/articles/dataset/Insider_Threat_Test_Dataset/12841247
[15] G. J. Silowash, D. M. Cappelli, A. P. Moore, R. F. Trzeciak, T. Shimeall, and L. Flynn, "Common sense guide to mitigating insider threats," 2012.
[16] U. Grupo en Redes y Servicios de Telecomunicación e Internet, "SPEDIA system for predition and early etection of insider attacks," https://spedia.lab.dit.upm.es/, consultado el 10 de marzo de 2025.
[17] M. ATT&CK, "Mitre att&ck," *URL: https://attack. mitre. org*, 2021.
[18] "Zenodo — zenodo.org," https://zenodo.org/, consultado el 19 de marzo de 2025.

Port2Vec: Generación de *embeddings* a partir de descripciones de servicios de red para su uso en Inteligencia Artificial aplicada a la Ciberseguridad

Eñaut Genua, Mikel Iturbe, Iñaki Garitano, Xabier Etxezarreta, Aitor Aguirre y Urko Zurutuza
Mondragon Unibertsitatea
Loramendi 4, 20500 Arrasate.Mondragon
enaut.genua@alumni.mondragon.edu, {miturbe, igaritano, xetxezarreta, aaguirre, uzurutuza}@mondragon.edu

***Resumen*—En este trabajo se presenta un enfoque novedoso para la representación de información relacionada con servicios de red, para su posterior uso en aplicaciones de aprendizaje automático orientadas a la seguridad en redes de comunicaciones. La mayoría de las propuestas de investigación orientadas a generar sistemas de detección de anomalías o sistemas de detección de intrusiones en redes que se han analizado, codifican el puerto destino y/o el puerto origen de flujos o de paquetes utilizando distintas estrategias, ya sea tratándolos como variables categóricas, numéricas, o utilizando técnicas de codificación *one-hot*. En este trabajo se propone la generación de *embeddings* a partir de las descripciones de cada puerto desarrolladas por IANA, enriquecidas con descripciones de los puertos obtenidas a través de IA Generativa. Esta metodología permite capturar el significado semántico subyacente de los servicios asociados a cada puerto, proporcionando representaciones más ricas y contextualizadas para modelos de aprendizaje automático.**

***Index Terms*—Ciberseguridad, detección de intrusiones, ingeniería de características, aprendizaje automático**

Tipo de contribución: *Investigación original (límite 8 páginas)*

I. Introducción

En el ámbito de la ciberseguridad y el análisis de tráfico de red, los números de puerto juegan un papel fundamental en la identificación de servicios y protocolos utilizados en la comunicación entre dispositivos. Tradicionalmente, los modelos de aprendizaje automático que trabajan con datos de tráfico de red, ya sea analizando flujos o paquetes de red, utilizan diferentes estrategias para codificar los números de puertos. En los casos en los que se usa esta característica como una variable numérica continua [1], [2], presupone erróneamente que dos números de puertos correlativos son más parecidos o cercanos a aquellos que no lo son. A modo ilustrativo, si se tomaran los números de puerto 80, 88, y 443 asignados (pero no obligatoriamente) a los servicios HTTP, autenticación de Kerberos y HTTP sobre TLS, respectivamente, el tratamiento de esta característica como variable numérica en algoritmos de agrupación indicaría que Kerberos y HTTP son más cercanos o parecidos que el propio HTTP y HTTPS. Si se tomaran las características como variables categóricas, provocaría que cada número de puerto se considerase como única referencia para un servicio de red, obviando que, para este caso, HTTP y HTTPS pueden estar relacionados en una misma interacción entre atacante y víctima. Finalmente, en otros casos se emplean técnicas como la codificación *one-hot* [3], [4] para representar cada puerto como una nueva característica en el conjunto de datos para su entrenamiento de forma similar al uso de variables categóricas. Sin embargo, genera vectores de alta dimensionalidad, especialmente cuando se trabaja con una gran cantidad de puertos (comportamientos relacionados con escaneos de red, denegación de servicio,...), provocando la necesidad de altas capacidades de computación y multiplicando el tiempo necesario para el entrenamiento y la inferencia de los modelos de aprendizaje.

Para abordar esta problemática, proponemos **Port2Vec** (todo el código está disponible en este repositorio[1]), un enfoque basado en técnicas de Procesamiento del Lenguaje Natural (PLN) para la generación de *embeddings* a partir de las descripciones textuales de los puertos de red. En lugar de tratar los números de puerto como valores discretos sin contexto, utilizamos técnicas de representación de texto para capturar la información semántica de los servicios asociados a cada puerto. Este enfoque permite que los modelos de aprendizaje automático aprovechen una representación más informativa y contextualizada, con la ventaja añadida de que los *embeddings* se computan una sola vez y permite una transcripción directa entre cada número de puerto y su representación vectorial.

En este trabajo, exploramos la metodología de generación de *embeddings* de descripciones de puertos, analizamos su efectividad en comparación con el uso tradicional de números de puerto y codificación *one-hot*. Los resultados sugieren que la incorporación de información semántica mediante *Port2Vec* puede mejorar la capacidad de los modelos para identificar relaciones y anomalías dentro del tráfico de red, y hacerlo en tiempos de computación razonables.

II. Fundamentos

El desarrollo de *Port2Vec* se sustenta en varios conceptos fundamentales del procesamiento del lenguaje natural y la representación vectorial de información, así como en el conocimiento establecido sobre puertos de red y sus servicios asociados.

II-A. Puertos de Red y Servicios

Los puertos de red son puntos de conexión lógicos que permiten la comunicación entre aplicaciones y servicios a través de redes TCP/IP. Cada puerto se identifica mediante un número entero sin signo de 16 bits, que debido a esto, hay $2^{16}-1$ puertos, y puede estar asociado a servicios específicos, como HTTP (puerto 80), SSH (puerto 22) o DNS (puerto 53). *Internet Assigned Numbers Authority* (IANA) [5] mantiene un

[1]Repositorio: https://gitlab.danz.eus/danz/segurtasuna/lip/port2vec.git

registro de asignaciones de puertos que incluye descripciones detalladas de los servicios y protocolos asociados a cada número de puerto. Estos puertos se asignan en varios modos, basándose en tres grupos principales:

- **Puertos de Sistema:** Los puertos de sistema, o puertos bien conocidos, abarcan el rango del puerto 0 al 1023, y están asignados a servicios fundamentales como HTTP, SSH o SMTP. Su función es ser el canal predecible para comunicaciones estándar. Estos puertos son controlados y asignados por IANA, y en la mayoría de los sistemas, el uso de estos puertos está restringido para los procesos con permisos privilegiados.
- **Puertos de usuario:** Los puertos de usuario, o puertos asignados, abarcan el rango del puerto 1024 al 49151 y son utilizados por programas que han sido oficialmente asignados para operar en un puerto específico, lo que permite tener una menor probabilidad de conflicto con otros servicios básicos. IANA no puede controlar el uso de estos puertos, pero sí que registra y asigna los usos como conveniencia a la comunidad.
- **Puertos dinámicos:** Los puertos dinámicos, también conocidos como puertos privados o efímeros, abarcan el rango del puerto 49152 al 65535, y nunca son ni serán asignados a ningún servicio por IANA, ni deberían por recomendación de los mismos.

A pesar de la existencia de estos grupos de puertos, nada impide que un servicio como SSH, que debería operar en el puerto 22, opere en otro puerto, como puede ser el 2222.

II-B. Representación Vectorial de Texto

La representación vectorial de texto, o *word embeddings* es una técnica fundamental en el PLN que permite convertir palabras o frases en vectores numéricos densos. Estos vectores capturan relaciones semánticas y sintácticas entre palabras, permitiendo que términos con significados similares se ubiquen cerca en el espacio vectorial. Modelos como *word2vec* [6] y *GloVe* [7] han demostrado la eficacia de estas representaciones en diversas tareas del PLN.

III. Trabajos Relacionados

El uso de representaciones vectoriales ha transformado la manera de abordar problemas tanto en el PLN como en el análisis de datos de red. En el ámbito del PLN, modelos como *Word2Vec* [6] y *GloVe* [7] han demostrado que es posible capturar relaciones semánticas entre palabras al proyectarlas en un espacio vectorial. La idea central es que elementos semánticamente similares queden próximos entre ellos en dicho espacio.

En el contexto de la ciberseguridad y el análisis de tráfico de red, ha sido común utilizar técnicas tradicionales como la codificación *one-hot* para representar atributos categóricos, incluidos los números de puerto o servicios. Este enfoque ha sido empleado en numerosos estudios de detección de intrusiones y clasificación de tráfico [3], [4]. Sin embargo, la codificación *one-hot* trata cada puerto como una categoría completamente independiente, lo que resulta en vectores de alta dimensionalidad, requiriendo mayores recursos y no captura ninguna relación semántica entre ellos. Para evitar eso, Figueiredo et al. [8] presenta una manera de evitar la gran dimensionalidad del vector resultante, capando el número máximo del puerto en 4096, y así, perdiendo la capacidad de distinguir los puertos mayores a 4096.

El uso directo de valores numéricos de puerto es un enfoque alternativo, empleado en diferentes estudios [1], [2], que también presenta limitaciones significativas. El uso directo de los números de puerto como características numéricas puede ser problemático, ya que la magnitud del número de puerto no tiene una relación directa con su función o comportamiento en la red. Por ejemplo, el puerto 21 (FTP) no está funcionalmente "más cerca" del puerto 22 (SSH) que del puerto 25 (SMTP), a pesar de su proximidad numérica.

De forma similar a este trabajo, Ring et al. presenta en [9] IP2Vec, como una medida de similitud para direcciones IP basada en Word2Vec. La propuesta aprende similitudes extrayendo información contextual de datos de red, considerando similares aquellas direcciones IP que aparecen en contextos similares. La evaluación experimental en dos conjuntos de datos públicos demuestra su efectividad para agrupar direcciones IP en redes botnet y, mediante métodos de visualización, confirma la capacidad de IP2Vec para capturar similitudes basadas en patrones de comunicación. La propuesta DarkVec y su posterior evolución i-DarkVec [10], [11] construyen un espacio de vectores o *embedding* único, estático e inmutable a partir de todos los datos observados en una darknet, sin ser específicas en las relaciones entre puertos.

En [12] se presenta un esquema de aprendizaje automático para rastrear actividades de ataque en la darknet mediante extracción y búsqueda de similitudes de características de puertos de destino con FastText, reducción dimensional con UMAP [13] y agrupamiento con DBSCAN [14]. Los experimentos utilizan tráfico de un sensor darknet /16 durante un mes para demostrar la eficacia del método propuesto. Esta investigación ya identifica una problemática similar, pero una aproximación que no se basa en convertir los puertos en características con semántica, sino en modelar los propios puertos para analizar las relaciones subyacentes.

Al igual que en nuestra propuesta, están emergiendo propuestas similares que ofrecen modelos pre-entrenados de tráfico para facilitar la labor de análisis. Peng et al. proponen recientemente PTU (*Pre-trained model for network Traffic Understanding*) en [15], como el diseño de un esquema de representación de tráfico que integra el contenido estático de los paquetes y las dinámicas de red en un espacio de entrada unificado. En esta ocasión también, los modelos y *embeddings* involucran todo el comportamiento del tráfico, no permitiendo identificar características intrínsecas de la red en un modelo de detección.

El presente trabajo introduce *Port2Vec*, como una metodología que genera *embeddings* de puertos a partir de sus descripciones textuales. Al igual que en PLN, esta estrategia permite que los puertos con significados o funciones similares—por ejemplo, el puerto 80 (HTTP) y 443 (HTTPS)—se representen mediante vectores cercanos en el espacio vectorial. Esta representación no solo reduce la dimensionalidad en comparación con la codificación *one-hot*, sino que también facilita la integración de la información semántica de los puertos en modelos de aprendizaje automático.

De este modo, *Port2Vec* se posiciona como una propuesta que aprovecha avances en técnicas de representación de características en espacios vectoriales (*embedding*) para superar

las limitaciones de las representaciones tradicionales y ofrecer una visión más rica y contextualizada de la funcionalidad de los puertos en entornos de red.

IV. Metodología

En este trabajo, se propone una metodología para generar representaciones vectoriales densas (*embeddings*) para el espacio completo de puertos de red.

Esta metodología transforma cada puerto en un vector de dimensión controlada que codifica sus características semánticas. El proceso (Figura 1) se estructura en cuatro fases secuenciales:

1. Adquisición de datos.
2. Preprocesamiento de datos sobre servicios de red.
3. Generación de representaciones vectoriales mediante modelos de *embeddings* neuronales.
4. Reducción de dimensionalidad preservando relaciones semánticas.

Adicionalmente, se implementa una variante aumentada utilizando *LLM*s (Modelos del Lenguaje de Gran Escala) para enriquecer el contenido semántico de los *embeddings*.

Un aspecto crítico de la metodología es su naturaleza agnóstica respecto a modelos específicos, permitiendo la sustitución de componentes individuales con implementaciones alternativas o modelos mejorados sin alterar el marco general. Esta característica garantiza la extensibilidad de la metodología de *Port2Vec* ante avances futuros en tecnologías de *embeddings* semánticos.

IV-A. Adquisición de Datos y Preprocesamiento

Como fuente de datos primarios se utilizó el registro oficial "*IANA Service Names and Port Numbers*" [5], que constituye la referencia autoritativa para la asignación global de puertos de red. Esta fuente proporciona metadatos estructurados para cada puerto.

La evaluación preliminar de los datos reveló heterogeneidad significativa en la completitud y calidad de los metadatos disponibles. Específicamente, se identificaron varios desafíos: ausencia de descripciones, falta de asignación de nombre de servicio, representaciones mediante rangos numéricos, y ausencia total de asignaciones para puertos dinámicos (49152-65535).

Para solventar dichas limitaciones, se implementó un método de procesamiento de los registros, estructurado en cuatro etapas:

1. **Normalización sintáctica:** Se aplicaron transformaciones determinísticas para estandarizar las representaciones textuales y numéricas, incluyendo resolución de valores nulos en protocolos mediante la asignación del designador "*any*", sustitución de nombres de servicio ausentes basado en patrones léxicos, y estandarización de variaciones ortográficas y sintácticas.
2. **Complementación de los datos:** Se añadieron entradas para los puertos sin representación explícita en el registro IANA, particularmente el segmento dinámico/privado (49152-65535), incluyendo la información asociada. Esta información asociada se ha metido manualmente, poniendo como protocolo el valor "*any*" y como descripción el valor "*Dynamic and/or Private Port*". Este procedimiento garantiza la exhaustividad del conjunto de datos, condición necesaria para la generación completa de *embeddings*.
3. **Expansión de rangos:** Se implementó una función de expansión que transforma rangos de puertos (e.g. "49152-65535") en registros individuales, generando entradas discretas para cada puerto mientras se preservan los metadatos asociados.
4. **Generación de características textuales:** Se diseñó un esquema de representación textual unificada que integra los atributos discretos en una descripción estructurada mediante una plantilla consistente que incluye nombre del servicio, número de puerto, protocolo, clasificación

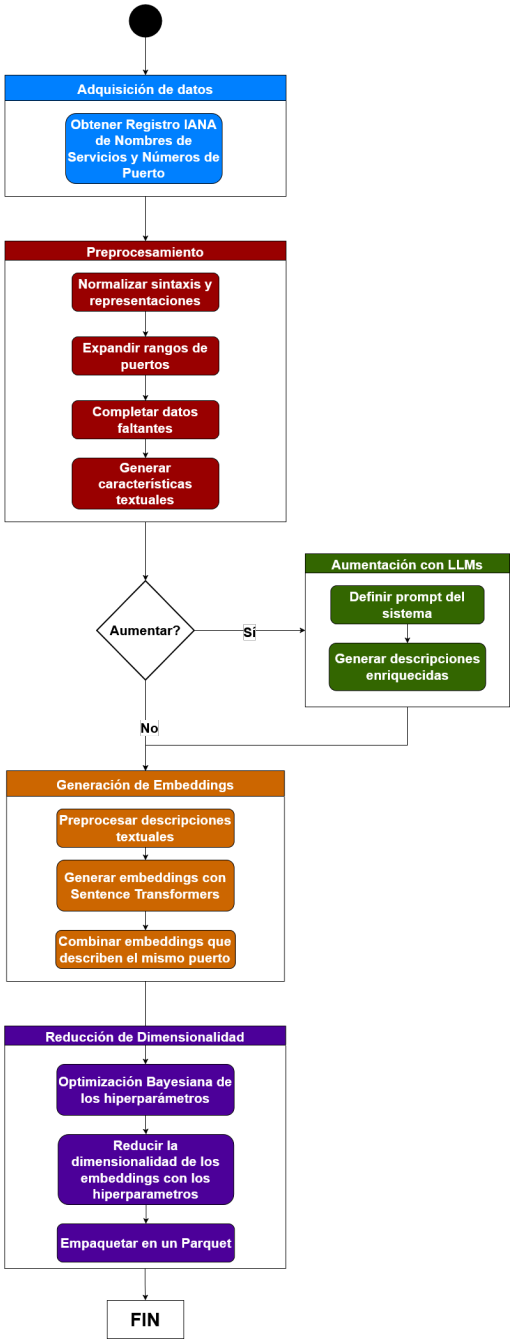


Figura 1. Diagrama de flujo de la metodología

(puerto de sistema/usuario/dinámico) y descripción original. Como ejemplo, una de las características textuales finales del servicio SSH es la siguiente: "*The 'ssh' service runs on network port 22 using the TCP protocol (system port). Description: The Secure Shell (SSH) Protocol*"

La verificación de la exhaustividad del conjunto de datos procesado se realizó mediante la comprobación de la representación de los 65.536 puertos posibles y la unicidad de cada puerto en el conjunto final.

IV-B. Aumentación con LLMs

Para enriquecer las características textuales del conjunto final, se implementó un procedimiento de aumento semántico utilizando modelos de lenguaje de gran escala (LLM). Este componente complementa la generación de características textuales mediante la incorporación de conocimiento técnico contextual no explícitamente presente en el registro de IANA, manteniendo la independencia del modelo específico utilizado. Por ejemplo, la descripción del puerto 153 (SGMP) en el registro dice: "*SGMP*", pero con el aumento semántico conseguimos: "*Simple Gateway Monitoring Protocol used for monitoring and managing network gateways*".

El procedimiento de aumento semántico sigue los siguientes pasos:

1. **Definición del *prompt* del sistema:** Se diseñó un *prompt* específico para guiar al modelo en la generación de descripciones técnicamente precisas y semánticamente enriquecidas, con directrices para preservar la información original mientras se añade contexto relevante.
2. **Generación de descripciones aumentadas:** Para cada puerto, se genera una descripción enriquecida utilizando la descripción original como contexto.

Aunque el modelo usado es *deepseek-ai/DeepSeek-R1-Distill-Llama-70B* [16], la arquitectura permite la sustitución del modelo de lenguaje utilizado para el aumento, adaptándose a la futura actualización a nuevos modelos con mayor capacidad de razonamiento técnico o conocimiento específico del dominio.

IV-C. Generación de Embeddings

La codificación de las representaciones textuales en vectores numéricos se realizó mediante un marco agnóstico de modelos, diseñado para funcionar con cualquier modelo de *embeddings* para frases, basándose en la tecnología de *Sentence Transformers* [17]. Esta aproximación desacopla la metodología de implementaciones específicas, permitiendo la integración fluida de avances futuros en modelos de *embeddings* sin modificaciones fundamentales al marco general. Esta flexibilidad es particularmente valiosa, considerando la rápida evolución del campo de los modelos de lenguaje y representaciones vectoriales.

El proceso de generación de *embeddings* sigue un flujo estandarizado que abstrae los detalles específicos del modelo subyacente:

1. **Preprocesamiento textual:** La descripción estructurada de cada puerto se estandariza en un formato común para todos los puertos, que incluye la información recolectada del conjunto de datos de IANA.



Figura 2. Proyección con UMAP [13] de los *embeddings* generados por *Linq-AI-Research/Linq-Embed-Mistral* [18] con aumentación.

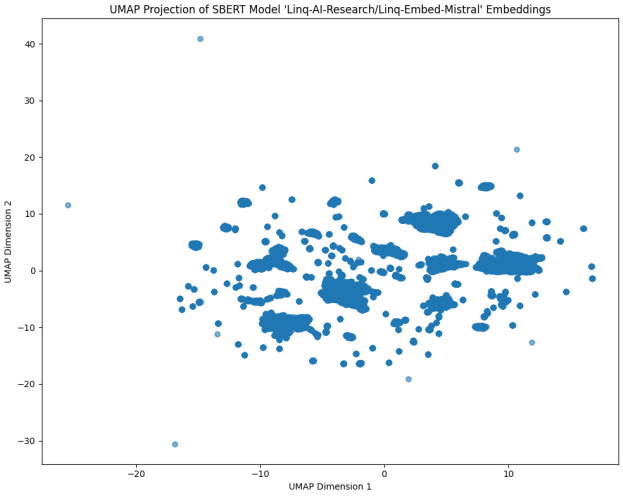


Figura 3. Proyección con UMAP [13] de los *embeddings* generados por *Linq-AI-Research/Linq-Embed-Mistral* [18] sin aumentación.

2. **Generación de *embeddings*:** Cada puerto se procesa mediante el modelo seleccionado para obtener su representación vectorial en la dimensionalidad nativa del modelo.
3. **Agregación para puertos multifuncionales:** Para puertos con múltiples entradas en el registro (debido a registros duplicados con distintos protocolos), se aplica una suma vectorial seguida de normalización para obtener una representación unificada.

Este procedimiento garantiza la consistencia en la generación de *embeddings* independientemente del modelo subyacente, facilitando la actualización futura con modelos mejorados.

El análisis de las visualizaciones UMAP muestra los beneficios del aumento mediante LLM en la generación de representaciones vectoriales de puertos de red. La comparación entre las dos proyecciones revela que las representaciones aumentadas por LLM (Figura 2) presentan agrupaciones notablemente más definidas que las representaciones no au-

mentadas (Figura 3). En la Figura 3, los puntos aparecen dispersos con límites difusos entre grupos, indicando una menor coherencia semántica donde puertos con funciones relacionadas pueden no estar consistentemente posicionados cerca entre sí. En contraste, la Figura 2 muestra clústeres claramente delimitados y concentrados, sugiriendo una organización más significativa basada en similitudes funcionales, donde puertos pertenecientes a servicios relacionados quedan agrupados coherentemente.

IV-D. Reducción de Dimensionalidad

Las representaciones nativas generadas por los modelos son vectores, siendo típicamente de alta dimensionalidad y son subóptimas para aplicaciones prácticas debido al coste computacional que pueden producir. Debido a esto, se implementó un marco agnóstico de técnicas de reducción dimensional, que permite la integración o sustitución de antedichas técnicas, adaptándose a diferentes requisitos de preservación topológica. Esta flexibilidad facilita la adaptación a diferentes casos de uso con requisitos específicos de preservación de estructura o eficiencia computacional.

La búsqueda de los mejores hiperparámetros para reducir la dimensionalidad de los *embeddings*, se realizó mediante un procedimiento bayesiano utilizando el algoritmo Tree-structured Parzen Estimator (TPE). Este enfoque modela la distribución de hiperparámetros condicionada al rendimiento observado, adaptando la estrategia de búsqueda según los resultados anteriores.

Un aspecto crítico de la metodología es la inclusión del algoritmo de reducción dimensional como un hiperparámetro dentro del espacio de búsqueda. Esto permite la selección automática del método más adecuado junto con sus parámetros específicos. Para cada técnica de reducción, se definió un subespacio de búsqueda de parámetros relevantes.

El procedimiento de optimización implementó varias estrategias para aumentar la eficiencia:

1. **Búsqueda inicial aleatoria:** Las primeras iteraciones exploran el espacio de hiperparámetros de manera aleatoria para construir un mapa inicial de los resultados.
2. **Búsqueda adaptativa:** Las iteraciones siguientes utilizan el conocimiento acumulado del mapa inicial para buscar hiperparámetros en regiones prometedoras del espacio de búsqueda.
3. **Detección automática de convergencia:** El proceso monitoriza la mejora relativa entre iteraciones, con un umbral de mejora mínima y un mecanismo de paciencia que detiene la optimización cuando no se observan mejoras significativas durante un número predeterminado de iteraciones.

IV-E. Evaluación de Calidad de Representación

La evaluación cuantitativa de la calidad de las representaciones se realizó mediante dos métricas complementarias que evalúan aspectos diferenciados de la preservación de la estructura:

1. ***Trustworthiness:*** Cuantifica la preservación de vecindades locales, evaluando en qué medida los puntos que son vecinos cercanos en el espacio reducido también lo son en el espacio original. Esta métrica penaliza las relaciones de vecindad "falsas" introducidas por la reducción, siendo una vecindad "falsa" aquella donde un punto aparece entre los k vecinos más cercanos en el espacio reducido pero no figura entre los k vecinos más cercanos en el espacio original. El parámetro k define el umbral que determina cuándo se considera que dos puntos son vecinos.
2. ***Shepard Goodness of Fit:*** Evalúa la preservación global de la estructura de distancias mediante el coeficiente de correlación entre las distancias por pares en ambos espacios.

Como métrica de distancia se usó la distancia del coseno, debido a que es la dirección lo que determina el carácter semántico del *embedding* y no la magnitud.

IV-F. Artefactos Resultantes

Los artefactos principales de este marco metodológico comprenden matrices de *embeddings* para cada combinación de modelo y dimensionalidad. Estos artefactos se generaron manteniendo la separación entre los derivados de las características textuales aumentadas y las originales.

Los *embeddings* finales se codificaron en formato Apache Parquet [19] con compresión Zstandard [20], estructurados como arrays multidimensionales de números de punto flotante con precisión de 32 bits, indexados por el número de puerto. Esta representación permite la fácil integración en otras aplicaciones de aprendizaje automático y análisis de redes.

La implementación final proporciona una API programática que facilita la lectura de los *embeddings* de los archivos Apache Parquet.

V. Entorno de Experimentación

La evaluación de los *embeddings* se realizó mediante un diseño experimental centrado en su aplicación práctica para tareas de clasificación de flujos de red. Este enfoque permite evaluar directamente la utilidad de los *embeddings* en escenarios reales de análisis de tráfico y seguridad.

En esta sección se describen las configuraciones y recursos utilizados para la evaluación experimental de los embeddings de puertos generados mediante Port2Vec. Se detallan tanto los aspectos técnicos del entorno computacional como las arquitecturas de aprendizaje automático empleadas para los experimentos de clasificación.

V-A. Infraestructura Computacional

Los experimentos se realizaron en una estación de trabajo con las siguientes especificaciones:

- CPU: AMD EPYC 7773X
- RAM: 130 GB
- GPU: NVIDIA RTX A6000 (48GB VRAM)
- SO: Ubuntu 22.04.5 LTS

Esta configuración permitió el procesamiento de los 65.535 puertos y la generación de *embeddings*, así como el entrenamiento paralelo de modelos de clasificación para la evaluación comparativa.

V-B. Conjunto de Datos para la Evaluación

Para esta evaluación se utilizó el conjunto de datos CIC-IDS-2017 mejorado [21]. Este conjunto proporciona flujos de red capturados durante varios días de actividad normal

y ataques simulados, con etiquetas detalladas para diferentes tipos de comportamiento malicioso.

El proceso de preprocesamiento fue exhaustivo para garantizar la calidad de los datos. Inicialmente, se eliminaron todos los registros que contenían valores nulos o presentaban inconsistencias que pudieran afectar el análisis. A continuación, se procedió con un filtrado riguroso para eliminar duplicados y valores atípicos extremos que podrían distorsionar los resultados. Los flujos clasificados como ataques intentados fueron reclasificados como benignos, ya que no representan amenazas reales al sistema.

Las variables categóricas presentes en el conjunto de datos fueron debidamente codificadas para su uso en los modelos de aprendizaje automático. Se realizó también una cuidadosa selección de características para reducir la dimensionalidad y mejorar la eficiencia del modelo. Un aspecto importante del preprocesamiento fue la eliminación deliberada de ciertos tipos de ataques, específicamente los escaneos de puertos y ataques de denegación de servicio (DoS), permitiendo así que el análisis se centrara en la detección de patrones de tráfico más sutiles y sofisticados.

Después de esto se generaron tres conjuntos de datos, y cada uno tenía una estrategia de preprocesamiento de los puertos diferente:

- Sin procesamiento especial (variable numérica continua).
- Codificación *one-hot*.
- Uso de los *embeddings* generados.

V-C. Arquitectura del Clasificador

Se implementó un clasificador de flujos basado en redes neuronales con una arquitectura diseñada para optimizar la detección de patrones en el tráfico de red. La estructura comienza con una capa de entrada adaptada específicamente a la dimensionalidad de las características extraídas durante el preprocesamiento. Para estabilizar el proceso de entrenamiento y acelerar la convergencia, se incorporó una capa de normalización por lotes inmediatamente después de la entrada.

El núcleo de la red está formado por cuatro capas ocultas densamente conectadas, con una configuración descendente de neuronas (512, 256, 128, 64) que permite una abstracción progresiva de los patrones del tráfico. Para preservar el carácter semántico de los *embeddings* a través de las diferentes capas de procesamiento, se implementaron conexiones de salto que mantienen la información esencial durante la propagación hacia adelante.

La robustez del modelo se mejoró mediante la incorporación de regularización por desactivación aleatoria de neuronas con una probabilidad de 0.2, reduciendo así el sobreajuste a los datos de entrenamiento. Finalmente, la arquitectura culmina con una capa de salida específicamente adaptada para la tarea de clasificación binaria, permitiendo discriminar eficientemente entre tráfico normal y malicioso.

V-D. Configuración del Entrenamiento

Para el entrenamiento del modelo se implementó una configuración robusta que permitiera capturar eficientemente los patrones distintivos del tráfico de red malicioso. Se seleccionó el optimizador *AdamW* [22] con una tasa de aprendizaje inicial de 0.01, aprovechando su capacidad para desacoplar la regularización del peso de los pasos de adaptación del gradiente, lo que resulta particularmente beneficioso en redes neuronales profundas.

La tasa de aprendizaje se gestionó dinámicamente mediante un planificador *CosineAnnealingWarmRestarts* [23], permitiendo ciclos de calentamiento que facilitan la exploración del espacio de parámetros y ayudan a escapar de mínimos locales subóptimos. Para abordar adecuadamente el desbalance inherente entre las clases de tráfico normal y malicioso, se implementó una función de pérdida *BCEWithLogitsLoss* con ponderación de clases, asegurando que ambas categorías recibieran la atención apropiada durante la optimización.

Los datos fueron distribuidos siguiendo un esquema estándar de división, asignando el 70 % al conjunto de entrenamiento, 15 % para validación y el 15 % restante para el test. El proceso de entrenamiento se estableció en 10 épocas y *PyTorch* [24] se configuró para que todas las operaciones fuesen reproducibles, para garantizar condiciones experimentales uniformes entre todas las estrategias.

VI. Resultados

Para la evaluación experimental se generaron *embeddings* en tres dimensionalidades: 16, 32 y 64 dimensiones. Estas configuraciones representan diferentes compromisos entre compacidad y capacidad representativa, permitiendo analizar el impacto de la dimensionalidad en el rendimiento. Aunque la metodología es agnóstica respecto al modelo, para estos experimentos se utilizó un único modelo de *embeddings*, *Linq-AI-Research/Linq-Embed-Mistral* [18], que se encuentra entre los modelos abiertos de mejor rendimiento según el benchmark MMTEB (*Massive Multilingual Text Embedding Benchmark*) [25] para inglés en HuggingFace. Se generaron con características textuales aumentadas y sin aumentar.

VI-A. Resultados de la Clasificación

Dada la naturaleza altamente desequilibrada del conjunto de datos, donde el tráfico legítimo supera significativamente al tráfico malicioso, se seleccionaron métricas específicas que proporcionan una evaluación robusta en escenarios de desequilibrio de clases:

- **F1-Score:** Representa la media armónica entre precisión y sensibilidad, proporcionando un equilibrio entre ambas. Esta métrica penaliza tanto los falsos positivos como los falsos negativos, siendo crucial en detección de amenazas donde tanto las falsas alarmas como los fallos de detección tiene costes significativos.
- **Average Precision:** Calcula el área bajo la curva precisión-sensibilidad, capturando el rendimiento del modelo a través de múltiples umbrales de decisión. Al ser independiente de un umbral específico, proporciona una evaluación más completa del rendimiento del clasificador en contextos desequilibrados.

Los resultados de clasificación para cada estrategia de representación de puerto se presentan en la Tabla I.

VI-B. Importancias de las Características

Para complementar el análisis del rendimiento de clasificación, se realizó un estudio detallado de la importancia de las características utilizadas por el modelo, con especial atención al impacto de los *embeddings* de puertos en la toma de decisiones del clasificador.

Tabla I
RENDIMIENTO DEL CLASIFICADOR DE FLUJOS CON DIFERENTES REPRESENTACIONES DE PUERTOS

Estrategia	***F1 Score***	***Average Precision***	***Precision***	***Recall***	**Tiempo de Entrenamiento**
Sin Procesar	0.92	0.81	0.89	0.96	17m 51.123s
One-hot	0.93	0.83	0.89	0.98	1h 37m 47.742s
Port2Vec 16D	0.96	0.95	0.94	0.98	17m 53.663s
Port2Vec 16D LLM	**0.98**	0.97	0.97	0.98	17m 57.751s
Port2Vec 32D	0.97	0.91	0.95	0.99	17m 58.595s
Port2Vec 32D LLM	0.96	**0.98**	0.93	0.98	17m 56.196s
Port2Vec 64D	0.97	0.97	0.95	0.98	17m 58.753s
Port2Vec 64D LLM	0.96	0.87	0.94	0.99	18m 4.568s

La evaluación de importancias se realizó mediante valores SHAP (*SHapley Additive exPlanations*) [26], una técnica fundamentada en la teoría de juegos cooperativos que permite cuantificar la contribución marginal de cada característica a la predicción final del modelo. Para el cálculo específico de las importancias, se siguió un proceso sistemático que permitió la comparabilidad entre las distintas estrategias de representación:

1. Se generaron muestras de fondo utilizando un subconjunto aleatorio del 0.5 % de los datos para cada estrategia.
2. Se seleccionó un conjunto de explicación correspondiente al 0.05 % de los datos para evaluar las contribuciones de las características.
3. Se calcularon los valores SHAP para cada característica en cada estrategia de representación.

En los casos donde los puertos se representan como vectores multidimensionales, la importancia se calculó mediante la suma de las importancias individuales de cada componente del mismo. Se utilizó la propiedad de *local accuracy* (precisión local), la cual asegura que la suma de las contribuciones individuales de las variables equivale a la predicción del modelo. Esto permite que, al descomponer una variable en varios subcomponentes, la suma de sus importancias refleje la importancia global de la variable original.

Los las importancias de los puertos para cada estrategia de representación de puerto se presentan en la Tabla II.

Tabla II
EL PORCENTAJE DE LA CONTRIBUCIÓN DE LOS PUERTOS

Estrategia	**Importancia de los Puertos (%)**
Sin Procesar	0.86 %
One-hot	21.73 %
Port2Vec 16D	45.45 %
Port2Vec 16D LLM	37.05 %
Port2Vec 32D	37.97 %
Port2Vec 32D LLM	41.84 %
Port2Vec 64D	51.96 %
Port2Vec 64D LLM	47.84 %

VII. CONCLUSIONES

Los resultados experimentales obtenidos en este estudio proporcionan evidencia sustancial sobre la efectividad de los *embeddings* de puertos de red para tareas de clasificación de flujos.

VII-A. Efectividad de los embeddings

Los *embeddings* demuestran un rendimiento superior en comparación con los enfoques tradicionales. Mientras que la representación directa sin procesamiento muestra un rendimiento base aceptable (F1: 0.92, AP: 0.81), queda claramente superada por los *embeddings* en todas sus variantes. Es de destacar que el *Average Precision* es la métrica que más se ve mejorada con el uso de los *embeddings*.

La codificación *one-hot*, a pesar de mostrar un rendimiento competitivo (F1: 0.93, AP: 0.83), conlleva una carga computacional excesiva debido a su alta dimensionalidad, como refleja su tiempo de procesamiento significativamente mayor.

VII-B. Impacto de la Dimensionalidad

El análisis de diferentes dimensionalidades revela patrones interesantes en el compromiso entre compacidad y capacidad representativa. Los resultados muestran que los *embeddings* de 16 dimensiones logran un rendimiento muy competitivo (F1: 0.96, AP: 0.95 para la versión estándar). Este hallazgo es particularmente relevante, pues demuestra que los *embeddings* muy compactos pueden capturar información suficiente para realizar eficazmente la tarea de clasificación de tráfico malicioso.

Al examinar configuraciones más amplias, se observa que los *embeddings* de 16 dimensiones alcanzan el rendimiento más alto del estudio, especialmente en su variante enriquecida con LLM, que consigue métricas casi perfectas (F1: 0.98, AP: 0.97). En los *embeddings* de 32 y 64 dimensiones se observa un rendimiento parecido al de los 16 dimensiones, incluso en la variante enriquecida con LLM para los de 32 dimensiones, pero en los de 64 dimensiones se puede observar un deterioro en las métricas. Esto puede indicar que el rendimiento puede ser impactado por la *maldición de la dimensión* en dimensiones altas, como también se puede ver en la variante de la codificación *one-hot*.

Estos resultados proporcionan una valiosa orientación práctica, indicando que en la selección de dimensionalidad se debe considerar cuidadosamente. El estudio demuestra que es posible obtener resultados altamente competitivos incluso con *embeddings* muy compactos, lo que podría ser crucial en entornos con limitaciones de recursos computacionales.

VII-C. Valor del Aumento Semántico Mediante LLM

La comparación entre las variantes estándar y aumentadas con LLM revela beneficios significativos del enriquecimiento semántico.

Las variantes aumentadas con LLM muestran consistentemente mayor *Average Precision* que sus contrapartes estándar, indicando una mejor capacidad para priorizar correctamente las instancias positivas. La mejora es particularmente notable en las representaciones de 32 dimensiones, donde el AP aumenta de 0.91 a 0.98, evidenciando que el conocimiento

contextual incorporado por el LLM proporciona información adicional. También es de destacar que en la variante de 64 dimensiones, el AP cae de 0.97 a 0.87, indicando que las altas dimensiones pueden hacer que el enriquecimiento semántico deteriore el rendimiento.

Estos resultados validan el valor del enfoque de aumento semántico, especialmente cuando se requiere alta precisión en la identificación de tráfico malicioso.

VII-D. Importancia de los Embeddings

El análisis de importancia de características revela hallazgos significativos sobre el valor informativo de los puertos en la clasificación. La Tabla II muestra que la representación sin procesar apenas contribuye a la toma de decisiones del modelo (0.86 %), mientras que la codificación *one-hot* mejora sustancialmente esta cifra (21.73 %).

Los *embeddings*, por otro lado, elevan considerablemente la importancia relativa de los puertos (37.05 %-51.96 %), indicando que capturan información semántica mucho más relevante para el modelo. Sin embargo, se observa que no existe una correlación directa entre la importancia del puerto y el rendimiento del clasificador. Los *embeddings* de 16 dimensiones aumentados con LLM alcanzan la mejor puntuación F1 (0.98) con una importancia relativa de 37.05 %, mientras que los de 32 dimensiones con LLM logran el máximo *Average Precision* (0.98) con 41.84 % de importancia.

Estos resultados evidencian que la calidad de los *embeddings* es más determinante que su mera contribución cuantitativa.

VII-E. Eficiencia computacional

La evaluación de eficiencia computacional presenta implicaciones prácticas importantes:

- La codificación one-hot resulta prohibitivamente costosa (1h 37m 47.742s) en el momento de entrenar el clasificador.
- Los *embeddings* ofrecen un balance superior entre rendimiento y eficiencia, con tiempos de procesamiento significativamente menores que la codificación one-hot.

Esta ventaja en eficiencia, combinada con su superior rendimiento predictivo, posiciona a Port2Vec como una solución práctica y efectiva para sistemas de análisis de tráfico de red y detección de intrusiones.

AGRADECIMIENTOS

Esta publicación es parte del proyecto de I+D+i PLEC2024-011222, financiado por AEI/10.13039/501100011033. Mikel Iturbe, Iñaki Garitano, Xabier Etxezarreta, Aitor Aguirre y Urko Zurutuza son miembros del Grupo de Investigación IT1676-22, financiado por el Departamento de Ciencia, Universidades e Innovación del Gobierno Vasco.

REFERENCIAS

[1] O. Belarbi, A. Khan *et al.*, "An intrusion detection system based on deep belief networks," in *International Conference on Science of Cyber Security*. Springer, 2022, pp. 377–392.

[2] D. Pujol-Perich, J. Suárez-Varela *et al.*, "Unveiling the potential of graph neural networks for robust intrusion detection," *ACM SIGMETRICS Performance Evaluation Review*, vol. 49, no. 4, pp. 111–117, 2022.

[3] M. Al-Qatf, Y. Lasheng *et al.*, "Deep learning approach combining sparse autoencoder with svm for network intrusion detection," *IEEE Access*, vol. 6, pp. 52 843–52 856, 2018.

[4] P. Wu and H. Guo, "Lunet: a deep neural network for network intrusion detection," in *2019 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2019, pp. 617–624.

[5] "Service Name and Transport Protocol Port Number Registry — iana.org," https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml.

[6] T. Mikolov, K. Chen *et al.*, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[7] J. Pennington, R. Socher *et al.*, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[8] J. Figueiredo, C. Serrão *et al.*, "Deep learning model transposition for network intrusion detection systems," *Electronics*, vol. 12, no. 2, 2023.

[9] M. Ring, A. Dallmann *et al.*, "Ip2vec: Learning similarities between ip addresses," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, pp. 657–666.

[10] L. Gioacchini, L. Vassio *et al.*, "Darkvec: automatic analysis of darknet traffic with word embeddings," in *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, ser. CoNEXT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 76–89.

[11] ——, "i-darkvec: Incremental embeddings for darknet traffic analysis," *ACM Trans. Internet Technol.*, vol. 23, no. 3, Aug. 2023. [Online]. Available: https://doi.org/10.1145/3595378

[12] S. Ishikawa, S. Ozawa *et al.*, "Port-piece embedding for darknet traffic features and clustering of scan attacks," in *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 593–603. [Online]. Available: https://doi.org/10.1007/978-3-030-63833-7_50

[13] L. McInnes, J. Healy *et al.*, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[14] M. Ester, H.-P. Kriegel *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[15] L. Peng, X. Xie *et al.*, "Ptu: Pre-trained model for network traffic understanding," in *2024 IEEE 32nd International Conference on Network Protocols (ICNP)*, 2024, pp. 1–12.

[16] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[18] J. Kim, S. Lee *et al.*, "Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement," Linq AI Research Blog, 2024. [Online]. Available: https://getlinq.com/blog/linq-embed-mistral/

[19] "Parquet — parquet.apache.org," https://parquet.apache.org/.

[20] Y. Collet and M. Kucherawy, "Zstandard Compression and the 'application/zstd' Media Type," RFC 8878, Feb. 2021. [Online]. Available: https://www.rfc-editor.org/info/rfc8878

[21] L. Liu, G. Engelen *et al.*, "Error prevalence in nids datasets: A case study on cic-ids-2017 and cse-cic-ids-2018," in *2022 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2022, pp. 254–262.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[23] ——, "Sgdr: Stochastic gradient descent with warm restarts," 2017. [Online]. Available: https://arxiv.org/abs/1608.03983

[24] J. Ansel, E. Yang *et al.*, "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, Apr. 2024. [Online]. Available: https://pytorch.org/assets/pytorch2-2.pdf

[25] K. Enevoldsen, I. Chung *et al.*, "MMTEB: Massive multilingual text embedding benchmark," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=zl3pfz4VCV

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774.

Verificación de edad en línea basada en extensión de navegador y pruebas de conocimiento cero

Alejandro Manuel López Gómez[1,2], Alonso Alfredo Rodríguez Rodríguez[2,4], Ofelia Tejerina Rodríguez[3], Julián Inza Aldaz[3,5], Gregorio López López[1,2]
[1]Instituto de Investigación Tecnológica (IIT), Universidad Pontificia Comillas
[2]Escuela Técnica Superior de Ingeniería (ICAI), Universidad Pontificia Comillas
[3]Observatorio LegalTech Garrigues – ICADE, Universidad Pontificia Comillas
[4]Polygon Labs
[5]EADTrust
amanlopezgomez@alu.icai.comillas.edu, alonso@polygon.technology, otejerina@comillas.edu, jinza@ext.comillas.edu, gllopez@icai.comillas.edu

Resumen- **La verificación de edad en línea se ha convertido en un desafío clave en la protección de menores y el cumplimiento normativo, a menudo entrando en conflicto con la privacidad del usuario. En este artículo, presentamos una solución basada en una extensión de navegador, tecnología *blockchain* y pruebas de conocimiento cero (ZKP, por sus siglas en inglés), que permite verificar la mayoría de edad sin revelar información personal sensible. Esta propuesta ofrece un equilibrio entre anonimato y usabilidad, beneficiando tanto a los usuarios, que pueden navegar con mayor privacidad, como a los proveedores de contenido adulto, permitiéndoles verificar la mayoría de edad sin almacenar datos innecesarios.**

***Index Terms*- blockchain, plug-in, verificación de edad, zero knowledge proofs**

Tipo de contribución: *Investigación original*

I. Introducción

En la era digital, la identidad en línea se ha convertido en un activo valioso, pero también en una vulnerabilidad. La recopilación masiva de datos por parte de gobiernos, empresas y ciberdelincuentes ha generado una creciente preocupación por la privacidad y la seguridad personal. Proteger la identidad digital no solo implica evitar fraudes o robos de datos, sino también preservar la libertad de expresión y el derecho al anonimato. Este anonimato es especialmente relevante en situaciones donde el estigma social o la discriminación pueden afectar la vida personal y profesional de un individuo. Por lo tanto, hay determinadas situaciones en las que sólo es necesario comprobar que el usuario cumple con un determinado requisito, sin necesidad de desvelar su identidad.

Un ejemplo claro es el acceso a contenido para adultos en línea, un sector que, a pesar de su legalidad en muchos países, sigue rodeado de tabúes y riesgos de exposición. La filtración de datos personales relacionados con este tipo de contenido puede dar lugar a situaciones de acoso o incluso represalias en entornos laborales y familiares. Por ello, el uso de herramientas de privacidad, como redes privadas virtuales (VPN), navegación anónima y métodos de pago discretos, se ha vuelto fundamental para quienes desean ejercer su derecho a la intimidad sin temor a consecuencias no deseadas.

En este contexto, el acceso a contenido para adultos en línea plantea un desafío importante tanto a nivel técnico como regulatorio: garantizar que solamente aquellos usuarios mayores de edad puedan consumirlo sin comprometer su privacidad. En muchos países, las regulaciones exigen mecanismos de verificación de edad, pero las soluciones actuales suelen ser invasivas, exigiendo el uso de documentos de identidad, tarjetas de crédito o registros en plataformas centralizadas. Esto no solo pone en riesgo la información personal de los usuarios, sino que también genera fricciones en la experiencia de navegación.

La raíz del problema radica en encontrar una solución que encuentre un equilibrio entre el cumplimiento normativo y garantizar la privacidad del usuario. Exigir datos personales para demostrar la mayoría de edad introduce riesgos como filtraciones de información, rastreo de actividad en línea y la pérdida de anonimato en un ámbito que, por su naturaleza, muchos usuarios prefieren mantener en privado. En este contexto, a nivel nacional ha surgido una solución técnica para la verificación de edad, la “Cartera Digital Beta”, que permite a los usuarios demostrar su mayoría de edad mediante credenciales digitales verificables.

En este artículo, se presenta una alternativa a esta solución basada en una extensión de navegador que emplea pruebas de conocimiento cero (*Zero Knowledge Proofs* - ZKP) para permitir la verificación de edad de manera anónima y eficiente. Posterior a la descripción de dicha solución se realizará una comparación con la “Cartera Digital Beta”, analizando las características de ambas soluciones en términos de privacidad, seguridad y usabilidad.

II. Contexto Legal

Los sistemas de verificación de edad (SVE) para limitar el acceso en línea a contenidos para adultos siguen siendo, desde hace años, un desafío recurrente en la protección de la infancia y la juventud. La solución no parece sencilla, ni siquiera única. Aunque en los últimos tiempos los esfuerzos por encontrar una propuesta suficientemente eficaz se han intensificado, la realidad es que las soluciones - técnicas y normativas - conocidas hasta la fecha, no acaban de establecer un marco sólido de confianza.

El legislador se debate entre la protección de los menores y la protección, al mismo tiempo, de los derechos y libertades de los consumidores adultos [1]. Por una parte, hay quien considera que los SVE pueden afectar a derechos y libertades de los propios menores, si los criterios que etiquetan contenidos "para adultos" son demasiado restrictivos o conservadores [2], también en las redes sociales que frecuentan, como sería el caso de Instragram, Tik Tok e incluso Onlyfans. Por otra parte, es altamente probable que los adultos renuncien a acceder determinados servicios o contenidos, como la pornografía o las apuestas en línea, si no confían suficientemente en la seguridad del sistema responsable del tratamiento, o en la correcta custodia de los datos identificativos que les solicitarían.

En Estados Unidos, Texas fue uno de los primeros Estados en promover legislación específica al respecto, la *Texas House Bill 896*, relativa a la "Prohibición del uso de plataformas de medios sociales por menores", de 2022, que siguieron Louisiana y Arkansas con medidas equivalentes. Para verificar la edad se exigía una fotografía del usuario y la licencia de conducir o similar, junto con la obligación del proveedor de contenidos de almacenarlo exclusivamente mientras procedía a la verificación y, una vez hecho, eliminarlo totalmente. La actual *Texas House Bill 1181*, relativa a la "Publicación o distribución de material sexual perjudicial para menores en un sitio web de Internet; sanción civil", de 2023, exime de responsabilidad en la verificación a los motores de búsqueda y a las plataformas de redes sociales, imponiéndosela a las empresas prestadoras del servicio para adultos si no verifican correctamente la edad del usuario, y esto causa daños a algún menor. Hoy son 19 los Estados que tienen leyes de verificación de edad parecidas y, por ejemplo, el sitio web Pornhub está bloqueado en 17 Estados [3]. Pero, "¿qué nivel de carga puede imponerse al acceso a la libertad de expresión en nombre de la protección de los niños frente al visionado de pornografía?". Esta cuestión ha sido llevada ante la Corte Suprema, *Free Speech Coalition et al. v. Paxton* [4], y se espera que sea resuelta en verano de 2025.

En Europa, la regulación de los SVE es un tema complejo que continúa avanzando sin una respuesta definitiva. Estos servicios están siendo desarrollados bajo varios marcos regulatorios, entre los que se incluyen el Reglamento (UE) 910/2014 (eIDAS), que establece las bases para la identificación electrónica y los servicios de confianza en transacciones electrónicas, y su revisión más reciente, el Reglamento (UE) 2024/1183 (eIDAS2). Este último introduce el Marco Europeo de Identidad Digital y la "Cartera de Identidad Digital" como método para la identificación en línea. Además, el Reglamento (UE) 2022/2065 (DSA) regula el mercado único de servicios digitales, modificando la Directiva 2000/31/CE.

Con el objetivo de mejorar la fiabilidad de los SVE, en enero de 2024 la Comisión Europea, a través de su Dirección General de Redes de Comunicación, Contenido y Tecnologías (Connect), creó un "Grupo de Trabajo sobre Verificación de la Edad". Este grupo tiene la misión de fomentar la cooperación con las autoridades nacionales de los Estados miembros que tienen experiencia en la verificación de la edad, para desarrollar prácticas más seguras y efectivas.

Por otro lado, en febrero de 2025, el Comité Europeo de Protección de Datos (CEPD), en colaboración con la Agencia Española de Protección de Datos (AEPD), adoptó un "Dictamen sobre la determinación de la edad para el uso de servicios en línea que requieren una edad mínima para acceder a ellos", basado en el Reglamento General de Protección de Datos (RGPD). Este dictamen establece principios clave para la implementación de los SVE, con el fin de garantizar que se respeten los derechos fundamentales de los usuarios.

Algunos Estados europeos han publicado sus propios criterios para el desarrollo de cualquier sistema de verificación, como Alemania (*Punktepapier einer gemeinsamen deutschen Position zum Thema Altersverifikation unter Inbezugnahme auf den DSA.* 31 de octubre de 2024, BfDI) o Reino Unido (*Age assurance for the Children's code*. 2024, ICO). Francia incluso ha aprobado legislación específica (*LOI n° 2023-566 du 7 juillet, 2023, visant à instaurer une majorité numérique et à lutter contre la haine en ligne*) y un detallado informe de especificaciones técnicas (*Référentiel déterminant les exigences techniques minimales applicables aux systèmes de vérification de l'âge mis en place pour l'accès à certains services de communication au public en ligne et aux plateformes de partage de vidéos qui mettent à disposition du public des contenus pornographiques*, 2024). En España, la *Ley 13/2022, de 7 de julio, General de Comunicación Audiovisual* obliga a las "plataformas de intercambio de vídeos a establecer sistemas de verificación de edad respecto a contenidos que puedan perjudicar a los menores" y es la Comisión Nacional de los Mercados y la Competencia (CNMC) la que evaluará la idoneidad de estos sistemas.

Entre otros, el Ministerio del Interior, el Ministerio de Juventud e Infancia, la CNMC, la Fábrica Nacional de Moneda y Timbre (FNMT) como entidad certificadora, y la AEPD, forman parte del Grupo de Trabajo interministerial que ha creado el Consejo de Ministros del Gobierno. Concretamente, la autoridad de control de protección de datos es la que ha venido liderando los trabajos de esta iniciativa pública, que se desarrollará e implementará, bajo el mandato del Ministerio para la Transformación Digital y de la Función Pública (MTDFP). En 2023 la AEPD publicó en su sitio web el "Decálogo de principios Verificación de edad y protección de personas menores de edad ante contenidos inadecuados" acompañado de una solución técnica que incorporaría en una aplicación móvil denominada "Cartera Digital Beta", y que promete cumplir su objetivo de verificar la edad del usuario sin afectar su privacidad. Este proyecto está además encuadrado en el marco legal que propone el "Anteproyecto de Ley Orgánica para la protección de las personas menores de edad en los entornos digitales" en el que trabaja en este momento el Gobierno español [5].

Si bien esta propuesta podría formar parte de la solución, porque plantea independizar la verificación de edad de los proveedores de identidad (no es preciso crear nuevos sistemas de identidad digital), fomentando así la neutralidad tecnológica y el libre mercado de opciones, con un sistema de claves públicas y privadas implementadas en una aplicación móvil y terceros de confianza intermediarios para la

verificación, no ha logrado todavía generar suficiente confianza en los usuarios. Lo cierto es que hasta la fecha ningún SVE se ha posicionado con fuerza en el mercado, ni se ha proclamado como "fiable" ni ha sido implementado en el sector público o privado [6].

III. Requisitos de los SVE

Una solución de verificación de edad en línea debe cumplir una serie de requisitos que garanticen el anonimato del usuario, así como la usabilidad del sistema tanto para éste como para el proveedor de contenido adulto. Estos requisitos han sido recopilados desde diferentes fuentes como el "Statement 1/2025 on Age Assurance" europeo [7] y Decálogo de Principios para la Verificación de Edad y Protección de Menores ante Contenidos Inadecuados publicado por la AEPD [8]. También se incluyen conclusiones propias fruto del desarrollo realizado.

El documento redactado por la AEPD establece diez principios fundamentales que todo sistema de verificación de edad debe cumplir desde un nivel de gestión, es decir, no se comentan posibles implementaciones o prácticas técnicas recomendadas. En primer lugar, el sistema debe evitar que se pueda realizar la identificación y rastreo de menores por parte de alguno de los actores involucrados en el proceso u otros terceros potencialmente maliciosos, previniendo así riesgos adicionales como la explotación o el abuso. La verificación de edad debe enfocarse en confirmar quién está autorizado a acceder a contenidos restringidos, sin exponer la identidad de los menores y sin ofrecer más información.

Otro principio clave es que la obligación de verificar la edad solo debe aplicarse a contenidos específicamente restringidos para adultos, evitando barreras innecesarias para el acceso a información general. Además, la verificación debe realizarse con métodos confiables y sin revelar datos sensibles, como la fecha de nacimiento del usuario. También se establece que todo sistema debe evitar el perfilado de usuarios basado en su navegación, así como la restricción de cualquier mecanismo que vincule la actividad de un usuario entre distintas plataformas para evitar su posible rastreo y posterior identificación

El decálogo de la AEPD también resalta la importancia de garantizar el derecho de los padres a decidir sobre los contenidos a los que acceden sus hijos, respetando la diversidad cultural y educativa de cada familia. Además, los sistemas de verificación de edad deben diseñarse de manera que respeten los derechos fundamentales, como la privacidad, la libertad de expresión y el acceso a la información. Por último, se subraya la necesidad de un marco de gobernanza claro, con mecanismos de supervisión y auditoría que garanticen la eficacia del sistema y la protección de los derechos de todos los usuarios.

Durante el desarrollo de la prueba de concepto, además de los requisitos impuestos por agencias como la AEPD, se han identificado requisitos técnicos que una solución de verificación de edad que busque su implantación a una escala mediana o elevada debe tener en cuenta. Desde el punto de vista de la privacidad del usuario, la solución debe garantizar la minimización de datos, es decir, se debe verificar la edad sin recopilar ni procesar información personal innecesaria. El único dato que debe validarse es la mayoría de edad, sin revelar otros detalles personales.

Además de otras regulaciones, es crucial que la solución cumpla con las normativas de privacidad, como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea, que exige el consentimiento explícito del usuario y la garantía y capacidad de destrucción los datos recopilados [9]. A nivel técnico esto quiere decir que ninguna solución podrá almacenar información acerca de sus usuarios, y aquella que fuese necesaria para su verificación debe ser destruida apropiadamente una vez se haya hecho uso de ésta.

En cuanto a la usabilidad de cara al usuario final, la solución debe ser fácil de instalar y utilizar. El proceso de verificación de edad debe ser lo más sencillo posible, sin requerir conocimientos técnicos. Debe poder completarse rápidamente, garantizando una experiencia de usuario fluida y sin fricciones. La verificación debe ser compatible con los principales navegadores y dispositivos, garantizando que cualquier usuario, independientemente de la plataforma que utilice, pueda acceder al contenido adulto sin dificultades.

Desde el punto de vista del proveedor de contenido, la solución debe facilitar el cumplimiento normativo. El proveedor debe poder verificar que los usuarios cumplen con el requisito de mayoría de edad sin necesidad de almacenar información personal sensible. La usabilidad para el proveedor de contenido es también esencial. La integración debe ser sencilla en las plataformas existentes, sin requerir cambios significativos en la infraestructura.

Respecto a su desarrollo, toda solución debe presentar un carácter modular y dinámico, ofreciendo la capacidad de incluir o retirar servicios sin afectar al funcionamiento del sistema. Por último, se debe ofrecer transparencia y auditoría, permitiendo que el proceso de verificación sea fácilmente auditado, aumentando la confianza en el sistema de las partes involucradas.

IV. Solución técnica propuesta

La solución ideada consta de tres componentes fundamentales (ver Fig. 1): el proveedor de contenido adulto, una extensión de navegador responsable de la interacción con el sistema y la generación de la prueba de conocimiento cero, y el propio sistema de verificación de edad y otorgamiento de credenciales. En primer lugar, el proveedor de contenido es responsable de bloquear el acceso al contenido adulto al usuario si no proporciona la prueba criptográfica pertinente, esto se ve ilustrado en los pasos 1 y 2 de la Fig. 1, donde el usuario trata de acceder al contenido y el servidor web realiza una petición para obtener su ZKP. En el paso 3, el usuario inicia el proceso para obtener su credencial y por tanto su prueba criptográfica. La verificación de la edad del usuario es realizada en el paso 4 (en la prueba de concepto desarrollada se ofrece verificación por certificado digital y/o teléfono móvil mediante una consulta al operador).

Habiendo sido exitosa esta verificación se prosigue al paso 5 donde se da la generación de credencial en el entorno de Privado ID, siendo enviada la credencial en el paso 6. La prueba es generada y almacenada por la extensión de navegador.

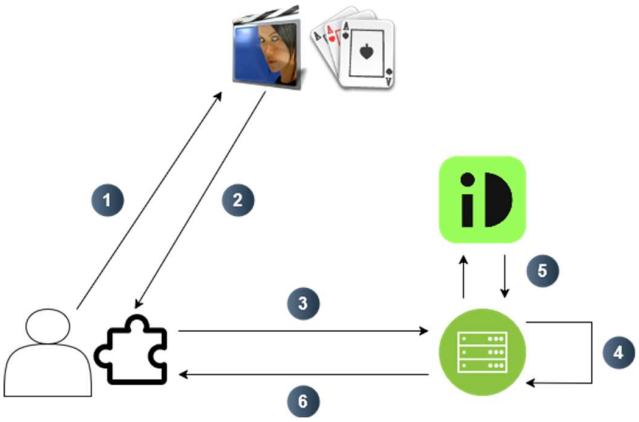


Fig. 1 Explicación a alto nivel del proceso de obtención de credencial

Se ofrece la opción de verificación de teléfono móvil debido a las siguientes razones. En España, algunos operadores como Movistar o Vodafone están empezando a implementar servicios de verificación de edad mediante el número de teléfono para garantizar que solo los usuarios mayores de edad accedan a ciertos contenidos o servicios.

Telefónica, a través de su iniciativa Open Gateway, ofrece la API estandarizada de Verificación de Número (Number Verification) [10].). Además, Telefónica ha colaborado recientemente con TikTok para desarrollar una solución que permite a los usuarios registrarse en la aplicación utilizando su número de teléfono. Esta iniciativa tiene como objetivo verificar la edad de los usuarios de manera eficiente, garantizando que solo aquellos que cumplen con los requisitos de edad puedan acceder a la plataforma.

Vodafone ofrece servicios de verificación de edad a través de su API *Age Verify*, diseñada para simplificar la autenticación de usuarios en productos restringidos por edad, como alcohol o tabaco [11|]. Esta API permite verificar si un cliente de Vodafone cumple con los requisitos de edad utilizando su número de teléfono móvil. El proceso se realiza de manera silenciosa a nivel de red y ofrece una confirmación binaria (Si o No), respetando la privacidad del usuario y ayudando a las empresas a cumplir con las normativas vigentes.

Cuando el usuario trate de volver a acceder al contenido (paso 7) esta vez se encontrará en posesión de una prueba de conocimiento cero, la cuál será enviada a la web adulta previa reclamación (paso 8). La web adulta le enviará esta prueba al entorno de Privado ID [12] (paso 9), en concreto al nodo de verificación o *"verifier node"*, realizándose una verificación de la prueba criptográfica *on-chain* u *off-chain* (este proceso se detalla más adelante). Tras recibir la confirmación de este nodo, la web adulta podrá emplear sus propios mecanismos de acceso (JWT, cookies…) para permitir al usuario acceder al contenido (ilustrado en el paso 10).

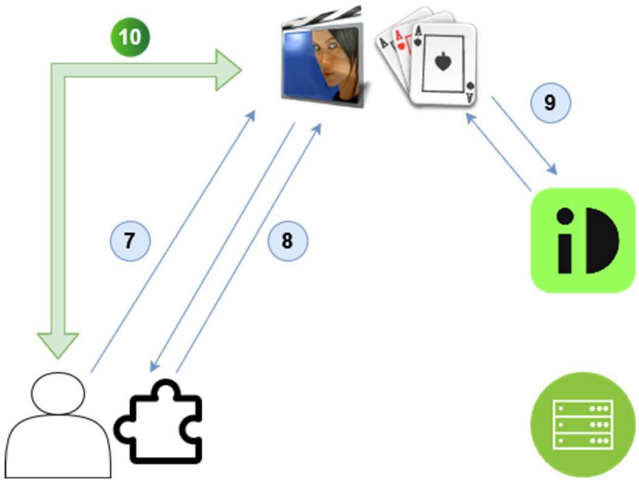


Fig. 2 Explicación a alto nivel del acceso a contenido adulto mediante ZKP

Privado ID es un sistema de identidad descentralizada que permite a los usuarios demostrar ciertos atributos sin revelar información sensible. Utiliza *Zero-Knowledge Proofs* (ZKP) y *blockchain* para garantizar la privacidad. Dentro de este ecosistema, hay dos componentes clave: el *issuer node* y el *verifier node*

El *issuer node* es la entidad encargada de emitir credenciales verificables cuando un usuario necesita demostrar cierta información, generando una credencial firmada digitalmente. El usuario recibe esta credencial y la almacena en su dispositivo. Esto permite que los usuarios tengan control total sobre sus credenciales y decidan cuándo y con quién compartirlas.

Por otro lado, el *verifier node* es el componente que valida las credenciales presentadas por los usuarios sin comprometer su privacidad. Cuando un servicio o aplicación (en este caso el proveedor de contenido adulto) necesita verificar cierta información, solicita una prueba al usuario. En lugar de compartir toda la credencial, el usuario genera una prueba criptográfica con ZKP, que permite al *verifier node* confirmar la autenticidad de la información sin revelar detalles adicionales. Esto garantiza que los servicios puedan verificar atributos específicos (en este caso si alguien es mayor de edad) sin darse a conocer más datos de los necesarios.

La interacción entre el usuario y el sistema se realiza a través de la extensión de navegador, la cual se comunica con el sistema, encargado de verificar la mayoría de edad del usuario. En las pruebas de concepto realizadas, este se ha constituido a partir de API REST *(Application Programming Interface - Representational State Transfer)* construidos en Python, y se han establecido dos posibles mecanismos de identificación, a partir de certificados digitales o mediante el número de teléfono, como ya se ha comentado anteriormente.

Una vez verificada la mayoría de edad por parte del sistema este realizará una serie de peticiones al *issuer node* del entorno de Privado ID. Este genera la credencial que será devuelta al usuario como una URL. La interacción por parte del usuario es necesaria debido a cómo está construido Privado ID (tras hacer *click* en la URL se muestran detalles de la credencial y el usuario debe aceptarla haciendo *click* en un

botón). En futuras versiones se devolverá esta credencial como un código QR, permitiendo al usuario utilizar una cartera móvil.

La extensión descrita anteriormente aparte de la lógica necesaria para la interacción con el resto del sistema presenta una cartera digital capaz de almacenar esta credencial. Esto se hará a partir de un fichero cifrado gestionado y almacenado por la extensión. Tras este proceso de verificación y asignación, en el momento en el que el usuario accede al sitio de contenido adulto, la extensión compartirá la prueba criptográfica al servicio web por medio de una petición que debe ir a través de HTTPS en el cuerpo o cabeceras de esta. El proveedor de contenido adulto contactará entonces al *Verifier node*. La verificación podrá realizarse *off-chain* u *on-chain.*

En el caso de verificación *off-chain* el proceso sería el siguiente. Al acceder al sitio web de contenido adulto, el usuario debe enviar una solicitud a la extensión (en el caso de la verificación web). Ésta envía la ZKPP generada previamente a partir de las credenciales guardadas en la propia cartera del plugin. El usuario entonces envía la ZKPP al *Verifier node* y este la verifica utilizando la API de verificación.

El *verifier node* verifica entonces que el estado del emisor de la credencial (el *issuer node* que otorgó esta credencial) y el estado del usuario sigan siendo válidos y no hayan sido revocados, ambas validaciones se realizan en segundo plano a partir de llamadas a dos contratos alojados en la red (se explica con mayor detalle en el siguiente párrafo). Si la verificación es exitosa, el *verifier node* devuelve un OK al proveedor de contenidos para que este emplee su lógica (cookies, JWT…) para otorgar y mantener por un periodo de tiempo el acceso del usuario al contenido.

Existen dos formas de agregar la lógica de verificación a ZKP a un contrato [13]. Se puede heredar la lógica del contrato *EmbeddedZKPVerifier* ya desplegado o emplear el contrato *UniversalVerifier*. Ambos contratos comparten la misma clase principal e implementan la misma interfaz *IZKPVerifier*, que define métodos para establecer, obtener y enviar respuestas para solicitudes de pruebas.

Mediante *EmbeddedZKPVerifier* se agrega la lógica de verificación al contrato personalizado del cliente, eso significa que todos los resultados de verificación se almacenan en el estado del contrato de este cliente, por tanto, si se quiere usar la verificación *on-chain* para un nuevo contrato, este debe volver a enviarle todas las respuestas a pruebas previamente enviadas.

El proceso para la verificación de una prueba es el siguiente. Previo a este proceso, se debe haber implementado un contrato personalizado de cliente y diseñar una solicitud de prueba con el método de contrato *setRequest*. Cuando un usuario accede al contenido web adulto, este se comunica con la extensión enviando una solicitud de verificación. El plugin de navegador entonces envía la ZKPP generada en la *wallet* al obtenerse la credencial. Esta se envía al proveedor de servicios y este la envía al contrato a través del método *submissionZKPResponse* o *submissionZKPResponseV2*.

El contrato entonces verifica que el estado del emisor de la credencial (el estado del *issuer node*) y el estado del usuario sigan siendo válidos y que no hayan sido revocados. Si la verificación es exitosa, se informa al sitio web para que este aplique su lógica de acceso para el usuario.

Por otro lado, *UniversalVerifier* se implementa como un contrato independiente y actúa como un registro central de verificación. Una vez que se envía una prueba, se puede usar en muchos contratos de cliente diferentes sin necesidad de enviar a estos las pruebas previas.

V. PRUEBAS DE CONOCIMIENTO CERO

Una prueba de conocimiento cero (ZKP) permite a un usuario o parte convencer a un verificador de la veracidad de una declaración sin revelar información adicional. Privado ID utiliza principalmente zk-SNARKs (Zero-Knowledge Succinct Non-Interactive Argument of Knowledge) a través del protocolo Iden3. Las zk-SNARKs generan pruebas de tamaño reducido y permiten una verificación rápida. El protocolo Iden3 implementa un esquema llamado Rapid-PLONK para una generación y verificación eficiente de zk-SNARKs [14]. Circom, un compilador de circuitos, utilizado para definir la lógica de las pruebas de conocimiento cero en Iden3. El protocolo Iden3 es de código abierto y se centra en la identidad auto-soberana, la privacidad y la descentralización. Proporciona un marco para crear y gestionar identidades digitales utilizando blockchain y pruebas de conocimiento cero [15].

Las pruebas de conocimiento cero (ZKPs) son esenciales para la privacidad y minimización de datos en Privado ID, permitiendo a los usuarios demostrar atributos sin revelar los datos subyacentes. Esto minimiza la cantidad de datos personales compartidos. Privado ID también implementa la divulgación selectiva, dando a los usuarios control sobre la información compartida.

VI. PRUEBA DE CONCEPTO

Para demostrar el funcionamiento de esta propuesta, se ha desarrollado una prueba de concepto ofreciendo para el proceso de verificación inicial dos alternativas: mediante el número de teléfono del usuario y consultando al operador, o a través de un certificado digital de la FNMT (Fábrica Nacional de Moneda y Timbre).

En el caso del proceso de verificación de edad basado en el número de teléfono, intervienen varios actores clave: el usuario, el plugin de navegador que gestiona la verificación, el proveedor web que almacena y ofrece el contenido adulto, varios módulos API REST construidos en Python encargados del manejo de solicitudes, el operador móvil encargado de verificar la edad del abonado y el *issuer node* de Privado ID responsable de generar la credencial del usuario.

El proceso (ilustrado en la Fig. 3) comienza cuando el usuario solicita acceso al contenido restringido. En este punto, el servidor web realiza una petición al plugin de navegador requiriendo su ZKP. Como en este caso no se ha realizado el proceso de verificación previamente, la extensión de navegador inicia el proceso de verificación, ofreciendo al

usuario dos opciones, verificar su edad mediante su número de teléfono o utilizando un certificado de la FNMT.

Una vez que el usuario elige la verificación con su número de teléfono y procede a su introducción en el formulario, se envía a la API de Python. Esta genera una contraseña de un solo uso (One Time Password - OTP) realizando una petición HTTPS incluyendo el número de teléfono al servicio *Twilio* [16]. El usuario debe introducir este código en el formulario pertinente para proceder a recibir su credencial.

Tras confirmarse que el OTP enviado por el usuario y el generado por *Twilio* coinciden, se realiza una petición a otra API de Python que simula la interacción con el operador. El operador procede a comprobar si el número de teléfono pertenece a una persona mayor de 18 años, devolviendo una confirmación binaria sin más información. Si el número cumple con este requisito, se genera una petición de credencial que es enviada al nodo emisor.

Finalmente, el nodo emisor responde con una URL de credencial que el usuario puede utilizar para demostrar que ha superado la verificación de edad sin necesidad de revelar más información personal.

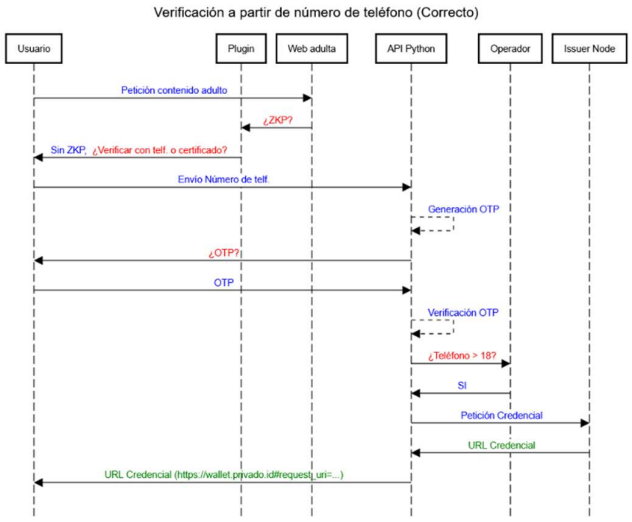


Fig. 3 Verificación a partir de número de teléfono

La Fig. 4 muestra el proceso de verificación de edad basado en un certificado digital, específicamente el certificado FNMT (Fábrica Nacional de Moneda y Timbre), aunque podría utilizarse otro. En este flujo participan varios actores clave: el usuario, el plugin que gestiona la verificación, la web que aloja el contenido para adultos, un servidor Nginx [17] que maneja el tráfico, una API en Python que procesa la verificación y un nodo emisor de credenciales.

El proceso de verificación a partir del certificado de usuario de la FNMT comienza de manera similar. En este caso el usuario elige la opción de certificado FNMT y hace clic en una URL que redirige al usuario hacia un servidor web nginx con el objetivo de enviar su certificado al sistema.

El servidor recibe el certificado del usuario y verifica su firma a partir de certificados raíz de la propia FNMT [18]. En caso de que se valide correctamente esta firma, el servidor envía el certificado a la API de Python, la cual verifica de nuevo la propia firma (evitando posible *bypass* del servidor), el estado del certificado en la CRL (Lista de Revocación de Certificados) y fecha de expiración.

Si el certificado es válido, automáticamente se demuestra que el usuario es mayor de edad, ya que la FNMT solamente otorga certificados de usuario a personas mayores de 18 años. La API entonces envía una petición de credencial al nodo emisor, este genera una URL de credencial que es enviada al usuario.

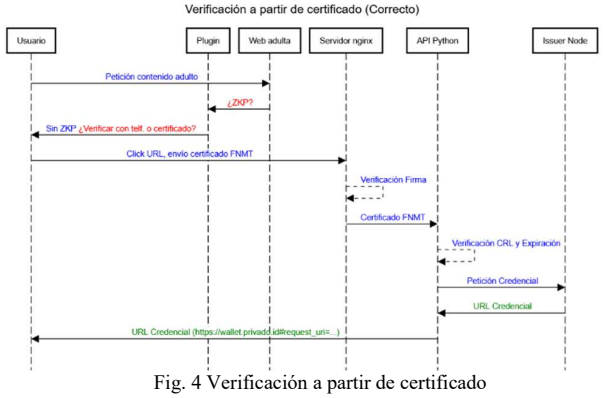


Fig. 4 Verificación a partir de certificado

El proceso de verificación de la edad del usuario finaliza cuando se envía a este la URL de su credencial. Cuando el usuario hace *click* en el *link*, se le redirige al entorno de Privado ID, donde entonces acepta la credencial y se produce la conexión con la cartera de identidad alojada en la extensión de navegador, como se muestra en la Fig. 5. La credencial es entonces almacenada y la cartera genera la ZKPP que será utilizada para la verificación con el proveedor de contenido adulto.

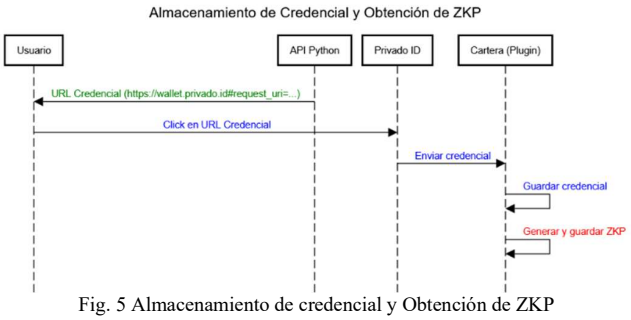


Fig. 5 Almacenamiento de credencial y Obtención de ZKP

Una vez obtenida la ZKP, cuando el usuario visite el sitio de contenido adulto, comienza la fase de verificación de la prueba. El flujo comienza cuando el usuario solicita acceso a dicho contenido. La web adulta, antes de conceder el acceso, envía una petición de verificación al *verifier node* para que este se comunique con la cartera del usuario.

El *verifier node*, al recibir la solicitud, pregunta a la cartera del usuario si dispone de una ZKP válida. En este punto el usuario ya ha realizado el proceso de verificación de edad, por tanto, ya posee dicha prueba, y esta se envía para su verificación. El *verifier node* reenvía la ZKP recibida al contrato de verificación ZKP, que se encarga de validar la prueba de manera descentralizada. Una vez verificada, el contrato responde confirmando que la prueba es correcta.

Después de validar la ZKP, el *verifier node* realiza una consulta adicional al contrato de estado para comprobar si el usuario o el *issuer node* han sido revocados. Si ninguno ha sido revocado, el contrato de estado responde con una confirmación de que todo sigue siendo válido. Con estas verificaciones completadas, el *verifier node* informa a la web adulta que el usuario cumple con los requisitos y que puede acceder al contenido. Finalmente, la web adulta concede el acceso mediante su propio mecanismo de acceso, completando así el proceso de verificación preservando en todo momento la privacidad del usuario.

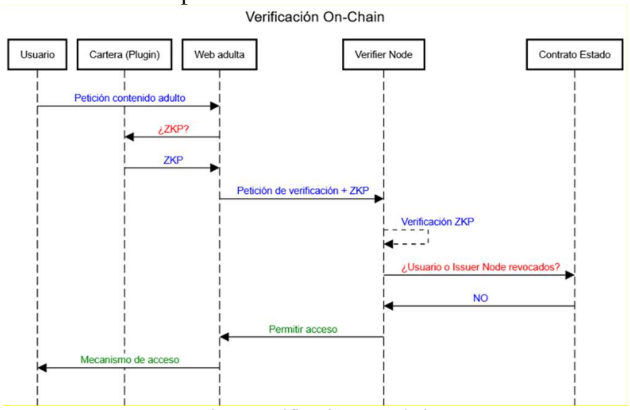


Fig. 6 Verificación On-Chain

Cabe destacar que durante el diseño de este último proceso (Fig. 6), se puso el foco en la facilidad de integración en la infraestructura del proveedor de contenidos adultos de esta solución, siempre considerando la dimensión de la seguridad. Debido a que la ZKP no contiene información alguna sobre la identidad del usuario, es seguro que el proveedor de contenido actúe como "proxy" entre la cartera del usuario y el nodo de verificación, evitando la comunicación directa entre este nodo y los usuarios del servicio, reduciendo la superficie de ataque.

Otra posible opción, de mayor facilidad para el proveedor de contenidos, pero aumentando la exposición a posibles vulnerabilidades es hacer que este redirigiese al usuario al nodo de verificación, realizase la verificación con este y el nodo devolviese al servicio web adulto un simple "apto/no apto". Finalmente se descartó esta opción en favor de la discutida en el párrafo anterior, ya que la complejidad técnica no aumenta en gran medida y se considera que se "compensa" esto con una mayor seguridad.

VII. Comparativa Entre Soluciones

La solución propuesta en este artículo se presenta como una alternativa a la opción propuesta por el Ministerio de Transformación Digital conocida como "Cartera Digital Beta". En este último apartado del documento se ofrece una comparación respecto a la seguridad de cada una. Es preciso destacar que ambas soluciones buscan resolver el mismo problema: permitir la verificación de mayoría de edad sin comprometer la privacidad del usuario.

La información mostrada en este artículo acerca de la solución propuesta por la administración pública ha sido obtenida de los documentos que esta misma ha puesto a disposición del público (véase [19]).

En términos de privacidad, la Cartera Digital Beta incorpora mecanismos de anonimización posteriores a la autenticación, se generan 30 pares de claves (clave pública y privada), junto con los DID respectivos, que le serán facilitados al emisor de las credenciales, de forma que cada credencial de mayoría de edad esté vinculada a una clave pública diferente. Además de la credencial de mayoría de edad, la solución "Cartera Digital Beta" será capaz de solicitar, almacenar y presentar las siguientes credenciales: padrón, ausencia de antecedentes por delitos sexuales, titulaciones universitarias y titulaciones no universitarias. No obstante, se mantiene un vínculo inicial con la identidad del usuario debido al uso de mecanismos como Cl@ve o DNI electrónico. Por el contrario, la solución propuesta refuerza la protección de la información personal mediante la aplicación de pruebas de conocimiento cero, que permiten validar atributos sin exponer los datos subyacentes, lo cual representa una mejora significativa en la preservación de la privacidad.

Respecto a la experiencia de usuario, la Cartera Digital Beta requiere el uso de una aplicación móvil y procesos de autenticación basados en credenciales gubernamentales, incluyendo la interacción mediante escaneo de códigos QR. La solución descentralizada, por su parte, plantea un flujo de uso más simplificado al operar desde una extensión de navegador y aprovechar métodos ya existentes, como los servicios ofrecidos por algunos operadores telefónicos y/o el hecho de que la FNMT emita certificados de usuario exclusivamente a usuarios mayores de edad, simplificando el proceso de verificación y ofreciendo una experiencia más fluida al usuario, lo que podría reducir la fricción tecnológica para el usuario final y el proveedor de contenido adulto.

Desde la perspectiva de seguridad, Cartera Digital Beta se beneficia de métodos consolidados y del respaldo institucional, aunque su centralización representa una vulnerabilidad estructural. En cambio, el enfoque descentralizado de la solución propuesta en este artículo mejora la resiliencia del sistema, disminuye la exposición de datos personales mediante pruebas de conocimiento cero, pero introduce riesgos emergentes vinculados al uso de contratos inteligentes, cuya seguridad depende en gran medida de una implementación rigurosa.

En lo que respecta a la escalabilidad, la propuesta centralizada se ve limitada por la infraestructura gubernamental disponible. La alternativa descentralizada enfrenta desafíos técnicos derivados de la eficiencia computacional de las pruebas criptográficas y de la capacidad operativa de la red, lo que puede afectar su viabilidad a gran escala.

Se destaca que ambas propuestas deben atender al cumplimiento normativo. La Cartera Digital Beta ha sido concebida en alineación con el marco regulador europeo eIDAS 2.0, mientras que la solución alternativa, aunque potencialmente compatible, deberá demostrar su conformidad con dicho reglamento, así como con la normativa de

protección de datos (GDPR), lo cual representa un reto significativo debido a la naturaleza distribuida del sistema.

En cuanto al potencial de integración, la propuesta gubernamental depende de mecanismos de control institucional como listas blancas y su implementación puede estar impulsada por mandatos legales. Por el contrario, la solución descentralizada presenta un modelo de integración más abierto, lo que facilita su adopción por parte de los proveedores de contenido adulto, incluidos proveedores independientes y plataformas emergentes.

Finalmente, los retos de adopción difieren sustancialmente. La Cartera Digital Beta requiere que los usuarios se familiaricen con herramientas gubernamentales y descarguen una aplicación específica, mientras que la alternativa basada en blockchain debe superar barreras de aceptación asociadas al uso generalizado de tecnologías de identidad digital descentralizada, aún incipientes en muchos contextos.

VIII. Conclusiones

Los sistemas de verificación de edad es solo una parte de un sistema de protección más amplio. Es una herramienta, así como la concienciación y formación acerca de los peligros del acceso no supervisado a las redes por parte de menores de edad. Para que esta sea efectiva, debe incluir un mecanismo confiable que no comprometa la privacidad de los usuarios y garantice su anonimato. La responsabilidad de aplicar estas medidas no solamente recae en los padres o tutores, es un deber social del que todos debemos ser conscientes. Se debe involucrar a todos los actores del ecosistema digital, incluyendo proveedores de contenido, motores de búsqueda, redes sociales y otros intermediarios digitales, con el objetivo de proteger a futuras generaciones de contenidos potencialmente dañinos.

Referencias

[1] Bacon, A. (2023). *Children's Safety in the Digital Age: A Look at Identity Verification Legislation in the United States*. American University (Washington, D.C.); Juris Mentem Law Review. https://doi.org/10.57912/23764731.v1.

[2] Stardust, Z., Obeid, A., McKee, A., & Angus, D. (2024). *Mandatory age verification for pornography access: Mandatory age verification for pornography access: Why it can't and won't 'save the children*. Big Data & Society, 11(2). https://doi.org/10.1177/20539517241252129

[3] Lovine, A. (2025, 2nd January). *All the states Pornhub is blocked in as of January 1. Mashable*. https://mashable.com/article/pornhub-shut-off-florida-12-other-states-8764866

[4] Mithani, J. (2025, 7th March). *Can states restrict content in the name of protecting children? The Supreme Court will weigh in*. Tech Policy Press. https://www.techpolicy.press/can-states-restrict--content-in-the-name-of-protecting-children-the-supreme-court-will-weigh-in/

[5] Pastor, J. (2025, 11th March). *Hace unos meses el Gobierno anunció el "pajaporte". Aún no sabemos nada de él pero las plataformas porno españolas sí. Xataca.* https://www.xataka.com/legislacion-y-derechos/gobierno-lleva-meses-retraso-lanzamiento-pajaporte-plataformas-porno-espanolas-estan-notando-su-impacto

[6] Jarvie, C., & Renaud, K.V. (2024). *Age Verification: Government Legislation, Supplier Responsibilization, and Public Perceptions. Children, 11*. https://doi.org/10.3390/children11091068

[7] Comité Europeo de Protección de Datos (2025). *Statement on Age Assurance*. https://www.edpb.europa.eu/system/files/2025-02/edpb_statement_20250211ageassurance_v1-1_en_0.pdf

[8] Agencia Española de Protección de Datos (s.f.). *Decálogo de principios de verificación de edad y protección de menores*. https://www.aepd.es/guias/decalogo-principios-verificacion-edad-proteccion-menores.pdf

[9] Diario Oficial de la Unión Europea (2016). *Reglamento General de Protección de Datos*. https://www.boe.es/doue/2016/119/L00001-00088.pdf

[10] Telefónica Open Gateway (s.f.). *API Number Verification*. https://opengateway.telefonica.com/actualidad/articulo/api-number-verification

[11] Vodafone Developer (s.f.). *Age Verify API Overview*. https://developer.vodafone.com/api-catalogue/age-verify/overview

[12] Privado ID (s.f.). *Privado Identity Verification*. https://www.privado.id/

[13] Privado ID (s.f.). *Verifier Overview*. https://docs.privado.id/docs/verifier/verifier-overview

[14] Protocolo Iden3. https://github.com/iden3

[15] Polygon ID Blog. https://polygon.technology/blog/introducing-polygon-id-zero-knowledge-own-your-identity-for-web3

[16] Twilio (s.f.). *Twilio Official Website*. https://www.twilio.com/en-us

[17] Nginx (s.f.). *Nginx Official Documentation*. https://nginx.org/

[18] Fábrica Nacional de Moneda y Timbre (s.f.). *Certificados raíz de la FNMT*. https://www.sede.fnmt.gob.es/descargas/certificados-raiz-de-la-fnmt

[19] Ministerio para la Transformación Digital (s.f.). *Especificaciones técnicas digitales*. https://digital.gob.es/especificaciones_tecnicas.html

.

Mejorando la seguridad de la cadena de suministro de software

Antonio Germán Márquez Trujillo, Ángel Jesús Varela-Vaca, María Teresa Gómez López, Antonia M. Reina Quintero, y Rafael Corchuelo
IDEA Research Group, Universidad de Sevilla
{amtrujillo, ajvarela, maytegomez, reinaqu, corchu}@us.es

***Resumen*—La seguridad en la cadena de suministro de software (SSC, por sus siglas en inglés) representa un desafío creciente debido a la complejidad de las dependencias y la gestión de sus vulnerabilidades. A medida que los proyectos de software incorporan múltiples componentes de terceros, se incrementan los riesgos de seguridad derivados de fallos de seguridad en estas dependencias o configuraciones inadecuadas. Este trabajo presenta un enfoque integral para abordar estos problemas mediante tres estrategias complementarias: (1) un modelo basado en el análisis de los inventarios de componentes y grafos de dependencias que permite la identificación y evaluación del impacto de vulnerabilidades en la SSC, (2) un proceso automatizado para la generación y enriquecimiento de informes sobre vulnerabilidades de una SSC, con el fin de mejorar la trazabilidad y priorización de riesgos, y (3) generar modelo de aprendizaje automático diseñado para predecir y recomendar exploits asociados a vulnerabilidades mejorando el testing de seguridad. Todas estas propuestas serán complementadas con herramientas para mejorar su integración en la toma de decisiones en ciberseguridad, optimizando la detección y gestión de amenazas dentro de los ecosistemas software y fortaleciendo la resiliencia ante ataques en la SSC.**

***Index Terms*—Cadena de Suministro de Software, Vulnerabilidades, Ciberseguridad, Desarrollo Software, Explotabilidad, Dependencias, Análisis Automático**

Tipo de contribución: *Investigación original*

I. Introducción

A día de hoy las cadenas de suministro de software (Software Supply Chain, SSC) juegan un papel fundamental en el desarrollo de software ya que permiten definir cómo se crean, distribuyen y mantienen las aplicaciones. Las SSC están compuestas por múltiples actores, como desarrolladores, proveedores de bibliotecas y frameworks, gestores de paquetes y plataformas de despliegue. Además, la creciente adopción de metodologías DevOps y la automatización de flujos de trabajo han convertido la SSC en un ecosistema dinámico y complejo [1].

Las SSC se sustentan en el uso de artefactos software de terceros también conocidos como dependencias [2], que permiten delegar funcionalidades en componentes externos. Sin embargo, esta dependencia de software introduce riesgos significativos de seguridad en el propio desarrollo. Para abordar estos riesgos, se han desarrollado enfoques como el Software Bill of Materials (SBOM), que permite definir un inventario detallado de los componentes utilizados en un proyecto de desarrollo, ya que facilita la identificación y trazabilidad de dependencias [3].

Como se ha comentado, el uso de dependencias externas expone a los proyectos a riesgos de seguridad , ya sea por uso de código con fallos, uso de dependencias obsoletas o la inyección de código malicioso en dependencias legítimas y ampliamente utilizados. Por tanto, es necesario incorporar capacidades de gestión de vulnerabilidades en las SSC [4], mediante un sistema que permita identificar los artefactos de software usados, analizar las vulnerabilidades que afectan estos artefactos y los exploits que podrían comprometer la seguridad del proyecto [5].

A pesar de los beneficios que aporta la SSC, su compleja superficie de ataque la convierte en un objetivo prioritario para ataques dirigidos. En los últimos años, se ha observado un aumento significativo en ataques a la SSC, como se demuestra en la inyección de código malicioso en paquetes legítimos [6], afectando a miles de proyectos en gestores de dependencias como Python Package Index (PyPI), Node Package Manager (NPM) y Maven Central (MVN) [7]. Estos ataques aprovechan la extensa red de dependencias de los proyectos de software, que incluyen tanto dependencias directas como indirectas [8], lo que complica la identificación, evaluación y mitigación de estas vulnerabilidades [9].

Además, la amplia variedad de versiones y configuraciones en los sistemas añade complejidad a la gestión de vulnerabilidades y la aplicación de estrategias de mitigación [10], [11]. Una única vulnerabilidad en un componente crítico puede comprometer toda la infraestructura de una SSC, permitiendo a actores malintencionados explotar fallos de seguridad [12].

Por otro lado, la gestión de vulnerabilidades no solo debe centrarse en la detección y mitigación de fallos de seguridad, sino también en garantizar la transparencia y trazabilidad de los componentes utilizados en un sistema. En este sentido, se han desarrollado estándares Vulnerability Exploitability eXchange (VEX) [13], que asociado al SBOM proporciona información específica sobre qué vulnerabilidades realmente afectan un producto o sistema en su contexto particular, ayudando a los equipos de seguridad a priorizar la mitigación de riesgos

Sin embargo, la generación y mantenimiento de estos reportes no está exenta de desafíos, ya que el desarrollo de un VEX requiere tiempo y recursos para analizar detalladamente qué vulnerabilidades representan una amenaza real en cada contexto específico. El número de vulnerabilidades documentadas no siempre refleja el riesgo real, dado que su impacto varía en función del contexto en el que se utilicen los componentes. Casos recientes han demostrado el alcance que pueden tener las vulnerabilidades en dependencias ampliamente utilizadas. Ataques como Log4Shell [14], que afectó múltiples versiones de la biblioteca Log4j en proyectos críticos [15], o la reciente CVE-2024-3094, que introdujo una puerta trasera en XZ Utils, utilizada en sistemas Linux, evidencian la necesidad de herramientas que permitan una evaluación más precisa del

riesgo.

A pesar de la creciente cantidad de vulnerabilidades documentadas, no todas son explotables de inmediato. De hecho, aunque existen más de 255,000 vulnerabilidades conocidas, solo un 9 % han sido asociadas con exploits registrados en bases de datos como VulnersDB [1]. Esto refuerza la importancia de herramientas como VEX, que permiten a las organizaciones diferenciar entre vulnerabilidades teóricas y amenazas reales, optimizando los esfuerzos en la gestión y automatización de la seguridad en proyectos de software.

Para abordar los problemas de la SSC y mejorar la gestión de vulnerabilidades, se han encontrado los siguientes desafíos clave a resolver. 1) **Análisis de las vulnerabilidades de dependencias en una SSC.** El primer desafío consiste en analizar el SBOM para identificar todas las dependencias de una SSC y evaluar su seguridad. Esto implica detectar vulnerabilidades conocidas en los componentes y determinar qué elementos representan un mayor riesgo dentro de la SSC. 2) **Mejorar la trazabilidad y transparencia sobre la seguridad de la SSC.** Una vez identificadas las vulnerabilidades, el siguiente desafío es la generación automática de un VEX y su versión enriquecida (eVEX), con el fin de proporcionar información precisa sobre la explotabilidad de cada vulnerabilidad en un contexto específico. Automatizar este proceso reduce la carga manual y mejora la trazabilidad de los riesgos en los proyectos de software. 3) **Mejorar el análisis de explotabilidad de vulnerabilidades.** Finalmente, el tercer desafío se centra en reducir el tiempo del análisis de explotabilidad de una vulnerabilidad, proporcionando recomendaciones optimizadas basadas en bases de datos de vulnerabilidades conocidas. Consiste en identificar el exploit más adecuado para cada vulnerabilidad sin exploit documentado, facilitando la evaluación del impacto y la priorización de medidas de mitigación para cada componente.

Para abordar los desafíos identificados en la SSC, proponemos diversas soluciones para mejorar la seguridad y optimizar gestión de vulnerabilidades en proyectos de software.

1. **Análisis de SBOM y las dependencias de proyecto software.** Para abordar el desafío del descubrimiento de vulnerabilidades en dependencias, hemos desarrollado una propuesta basada en grafos de dependencias, que permite modelar las relaciones entre paquetes y sus versiones dentro de la SSC. Este enfoque facilita la detección de vulnerabilidades en dependencias directas e indirectas, permitiendo evaluar su propagación y su impacto en la seguridad del software. Para validar su eficacia, hemos aplicado este modelo a cientos de repositorios públicos, comparándolo con soluciones comerciales y midiendo su precisión en la identificación de riesgos.
2. **Automatización en la generación y enriquecimiento de VEX.** Con el fin de reducir la carga manual en la gestión de vulnerabilidades en un SSC y así mejorar su trazabilidad y transparencia, hemos desarrollado una propuesta que permite la generación automática de VEX, integrando información enriquecida para mejorar su utilidad. Además, esta solución permite priorizar vulnerabilidades en función de su explotabilidad real dentro del software, proporcionando utilidad a los equipos de seguridad para tomar decisiones más precisas. La propuesta ha sido evaluada en repositorios con SBOM, midiendo su capacidad para mejorar la trazabilidad de vulnerabilidades y la eficiencia en la mitigación de riesgos.
3. **Recomendación de exploits para vulnerabilidades.** En la evaluación del impacto de una vulnerabilidad, es fundamental conocer si existen exploits efectivos o que afecten al proyecto y que permitan su explotabilidad. Para abordar este desafío, proponemos usar un modelo de recomendación, que analiza vulnerabilidades previamente documentadas y sugiere el exploit más adecuado dependiendo del contexto. Este enfoque permite reducir el tiempo de análisis y priorización de amenazas, facilitando la toma de decisiones en la gestión de seguridad. La validación de esta propuesta se ha realizado utilizando bases de datos de exploits, evaluando su precisión en distintos contextos y escenarios de uso.

Para el resto del artículo se va a seguir la siguiente estructura. En la Sección II, se van a detallar todas las soluciones desarrolladas para los desafíos mencionados en la introducción. En la Sección III, se comentan los resultados experimentales obtenidos para cada una de las propuestas. En la Sección IV, se recogen todos los trabajos relacionados de alta importancia para cada una de nuestras propuestas. Por último, en la Sección V, se darán unas breves conclusiones y trabajos futuros.

II. Soluciones Propuestas

Como podemos observar en la Figura 1, en una SSC, un proyecto en desarrollo usa dependencias directas e indirectas relacionadas. Las dependencias se especifican en el SBOM asociado al proyecto. Ciertas dependencias (directas o indirectas) de la SSC pueden contener vulnerabilidades, el uso de estas dependencias en el código del software en desarrollo hace que las vulnerabilidades heredadas de estas dependencias se introduzcan en el código fuente. Por lo que un tercero malicioso podría intentar explotarlas mediante la interfaz o API de los sistemas resultantes aumentando así la superficie de ataque.

A continuación, en la Subsección II-A, se detalla la solución propuesta para hacer frente al desafío sobre el análisis de SBOM y el modelado de dependencias en la SSC. En la Subsección II-B, se explica la solución propuesta para abordar la generación automática y enriquecimiento de un VEX. Por último, en la Subsección II-C, se describe la propuesta que diseña un modelo pequeño de lenguaje (SLM, por sus siglas en inglés) para la clasificación y recomendación de vulnerabilidades.

II-A. Análisis de SBOM y Modelado de Dependencias en la SSC

Uno de los principales desafíos en la SSC es la identificación y análisis de dependencias, tanto directas como indirectas [16]. Las vulnerabilidades pueden propagarse a través de múltiples niveles en la SSC, lo que dificulta la evaluación de su impacto real en el software. Sin una representación clara de la estructura de dependencias y sus versiones, los equipos de seguridad no pueden determinar qué componentes son

[1] Hasta el 29 de junio de 2024

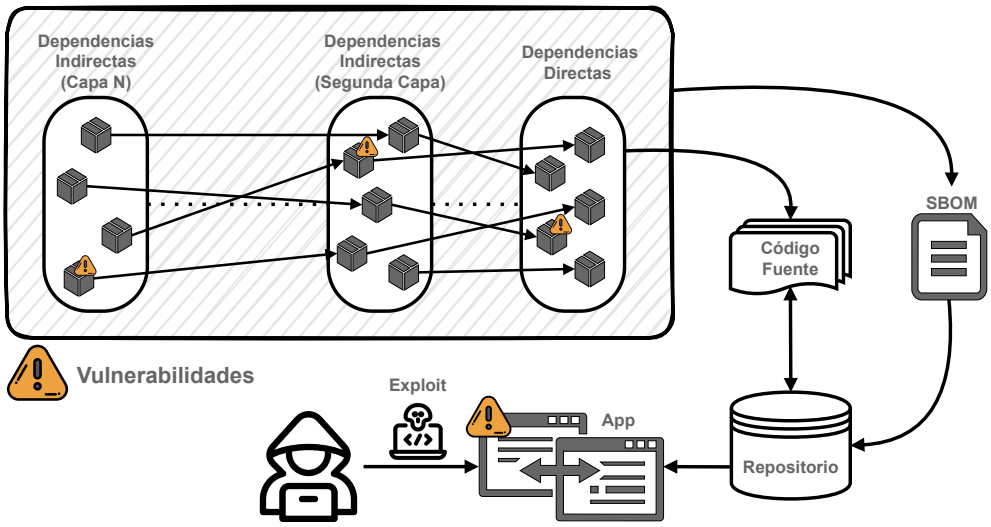


Figura 1. Superficie de ataque en la cadena de suministro de software.

críticos ni priorizar adecuadamente la mitigación de riesgos. Para abordar este problema, proponemos un modelo basado en grafos de dependencias atribuidos con información de seguridad. Este enfoque permite lo siguiente:

1. Extracción y modelado de la estructura completa de dependencias a partir del SBOM, incluyendo relaciones entre paquetes y versiones.
2. Atribución de vulnerabilidades documentadas en bases de datos de seguridad (la National Vulnerability Database (NVD) y la Open Source Vulnerabilities (OSV)) a versiones específicas dentro del ecosistema software.
3. Aplicación de técnicas de razonamiento para evaluar configuraciones seguras y optimizar la selección de versiones con menor impacto en seguridad.

A continuación, se detallan sus procesos clave.

II-A1. Extracción del Grafo de Dependencias a partir del SBOM: El primer paso en el análisis de la SSC es la construcción de un grafo de dependencias a partir de un repositorio de software. Para ello, se sigue un procedimiento en varias etapas, primero se identifican las dependencias directas a partir del Software Bill Of Materials (SBOM) o de los archivos de requisitos que dependiendo de la naturaleza del proyecto podrían ser un fichero *package.json*, *requirements.txt*, o *pom.xml*. Segundo, se crean nodos para cada dependencia con sus versiones usando la información de los repositorios oficiales, por ejemplo Maven Central o PyPI. Tercero, se establecen las relaciones entre dependencias filtrando versiones según las restricciones de compatibilidad. Y finalmente, se realizan los pasos anteriores recursivamente para dependencias indirectas.

Si una dependencia ya ha sido analizada previamente, se reutiliza la información para evitar extracciones redundantes, reduciendo el uso de API externas. El grafo resultante proporciona una base estructurada para el análisis de seguridad, permitiendo evaluar la conectividad del ecosistema de paquetes y detectar relaciones potencialmente vulnerables.

II-A2. Atribución de Vulnerabilidades: El grafo de dependencias se enriquece con información sobre vulnerabilidades, obtenida de los volcados públicos de OSV. La OSV contiene información de vulnerabilidades propias, vulnerabilidades que pertenecen a bases de datos de gestores de paquetes como la PYSEC para PyPI, vulnerabilidades de la GitHub Advisory Database, o incluso las vulnerabilidades de NVD, lo que aporta un amplio espectro de tipos de vulnerabilidades. Las vulnerabilidades se identifican mediante los identificadores únicos propios que tienen (CVE-ID, GHSA-ID, OSV-ID) y se asocian a versiones específicas de paquetes utilizando las enumeraciones CPE correspondientes. Esta información permite evaluar el impacto de cada versión en función de su severidad según el Common Vulnerability Scoring System (CVSS), almacenando estos valores para su posterior análisis y facilitando la toma de decisiones para la mitigación de riesgos de seguridad.

II-A3. Razonamiento sobre grafo de dependencias: El grafo generado se transforma en un modelo formal basado en Satisfiability Modulo Theories (SMT), permitiendo el análisis de configuraciones óptimas. Aunque este enfoque implica el uso de técnicas formales avanzadas, puede entenderse como una forma sistemática de buscar combinaciones de dependencias con el menor riesgo posible. Se utiliza el solver SMT, en nuestro caso Z3, para convertir las dependencias y versiones en variables booleanas o enteras, donde cada versión tiene un impacto calculado a partir de su puntuación CVSS. Esto permite realizar inferencias sobre qué combinaciones de dependencias ofrecen el menor riesgo de seguridad. Con el grafo de dependencias transformado a SMT, se pueden aplicar las siguientes operaciones de análisis:

- **Información estructural del grafo:** Permite conocer el número total de dependencias directas e indirectas, relaciones entre ellas y número de vulnerabilidades que las afectan.
- **Validez del modelo:** Determina si existe una configuración de versiones del proyecto que cumpla con todas las restricciones de compatibilidad y seguridad.
- **Optimización del impacto de vulnerabilidades:** Permite encontrar la configuración de dependencias que op-

timizan (minimizan o maximizan) el riesgo de seguridad.

- **Filtrado de configuraciones por criterios de seguridad:** Se pueden establecer umbrales de riesgo para seleccionar únicamente versiones de paquetes con impacto aceptable, lo que permite ajustar la configuración a la postura de seguridad que requiera el desarrollo.
- **Recomendación de configuraciones seguras:** La propuesta puede completar configuraciones parciales, proporcionar configuraciones por un impacto definido, o incluso expresar si una configuración es válida.

Para optimizar el rendimiento, las operaciones que generan múltiples configuraciones pueden limitar la cantidad de resultados devueltos. Además, el impacto de seguridad se puede calcular con diferentes métodos, como media o media ponderada, permitiendo flexibilidad en el análisis.

A partir de este modelo, se ha desarrollado una implementación denominada Depex[2], que permite analizar grandes ecosistemas de software y aplicar razonamiento sobre ellos de manera automatizada.

II-B. Automatización en la Generación y Enriquecimiento de VEX

La identificación y gestión de vulnerabilidades en la SSC es un proceso complejo, ya que no todas las vulnerabilidades documentadas en bases de datos de seguridad afectan realmente a un sistema. Muchas de estas vulnerabilidades dependen de cómo se utilizan los componentes de terceros dentro del software, lo que hace que la evaluación manual de su impacto sea una tarea costosa y propensa a errores. Para abordar este problema, se han desarrollado formatos como el VEX, que permiten documentar si una vulnerabilidad es realmente explotable en un sistema específico. Sin embargo, la generación manual de archivos VEX puede ser un proceso tedioso y difícil de mantener para una SSC compleja.

Si bien el proceso de generación de archivos VEX y eVEX implica varios pasos técnicos, puede entenderse como una forma estructurada de expresar si una vulnerabilidad realmente impacta el software en un caso concreto. Para mejorar este proceso, se propone un método automatizado para la generación y enriquecimiento de archivos VEX, basado en el análisis del SBOM, y mediante minería de repositorios de código y evaluación del código fuente. La solución propuesta permite determinar automáticamente si una vulnerabilidad en un componente de terceros tiene un impacto real en el software analizado. Se utilizan técnicas de análisis estático y dinámico, combinadas con información de bases de datos como NVD, OSV y GitHub Advisory Database, para inferir si un componente vulnerable está presente en el código y si sus funciones inseguras son utilizadas.

Dado que en muchos casos no es posible establecer automáticamente si una vulnerabilidad es explotable, hemos propuesto desarrollar un modelo complementario o modelo VEX extendido que llamaremos eVEX, que complementa el VEX con información adicional relevante para ayudar en la toma de decisiones. El eVEX incorpora detalles como la puntuación de impacto CVSS de la vulnerabilidad, identificadores CVE y CWE (Common Weakness Enumeration) de MITRE, existencia de código alcanzable (dentro del código fuente) y presencia de exploits en bases de datos públicas. Esto permite a los desarrolladores y equipos de seguridad la posibilidad revisar manualmente los casos en los que no se pudo inferir una conclusión automática, proporcionando un contexto más detallado sobre la vulnerabilidad y su posible explotabilidad.

Además, nuestra propuesta incluye un mecanismo de priorización de vulnerabilidades basado en factores de riesgo clave. La severidad de cada vulnerabilidad se calcula utilizando múltiples métricas, como la puntuación CVSS, la presencia de código alcanzable en el software y la existencia de exploits conocidos en bases de datos como Vulners y Packet Storm. Esta priorización permite a los equipos de seguridad enfocarse en los riesgos más críticos, optimizando la gestión de vulnerabilidades y reduciendo el tiempo necesario para responder a amenazas potenciales.

La solución propuesta automatiza gran parte del proceso de evaluación de vulnerabilidades, reduciendo la carga de trabajo de los equipos de seguridad y permitiendo una gestión más eficiente de los riesgos en proyectos que dependen de múltiples componentes de terceros. Al generar archivos VEX y eVEX de forma automática, las organizaciones pueden mejorar la trazabilidad de las vulnerabilidades dentro de la SSC y garantizar que los esfuerzos de mitigación se centren en las amenazas que representan un peligro real para sus sistemas.

Para demostrar la efectividad de este enfoque, está en desarrollo una implementación llamada VEXGEn[3], que genera automáticamente archivos VEX y eVEX a partir del análisis del SBOM, código fuente y bases de datos de seguridad.

II-C. Recomendación de Exploits para Vulnerabilidades

En la cadena de SSC, la explotabilidad de vulnerabilidades representa un desafío crítico en la evaluación de riesgos. Aunque existen bases de datos de seguridad que documentan vulnerabilidades conocidas, muchas de ellas carecen de exploits registrados, lo que dificulta la priorización de amenazas. No todas las vulnerabilidades tienen el mismo impacto; algunas pueden ser altamente explotables mientras que otras, aunque documentadas, no representan un riesgo real. La falta de información sobre su explotabilidad complica la toma de decisiones y la asignación de recursos para su mitigación.

Para abordar este problema, proponemos un modelo basado en un SLM que permite predecir la existencia de exploits potenciales para vulnerabilidades documentadas. En lugar de depender exclusivamente de bases de datos de exploits conocidos, este modelo analiza los atributos de las vulnerabilidades y establece patrones que permiten inferir la probabilidad de explotabilidad. De esta manera, se facilita la identificación de amenazas reales y se optimiza la priorización de vulnerabilidades en la SSC.

En esencia, el modelo busca responder a una pregunta práctica: ¿cual es el exploit funcional para esta vulnerabilidad? Para ello, se utilizan técnicas de aprendizaje profundo sobre atributos conocidos. El modelo se basa en la construcción de un conjunto de datos tabular estructurado, donde cada vulnerabilidad se representa mediante un conjunto de atributos clave, como vectores de ataque, plataformas afectadas y

[2]https://github.com/GermanMT/depex

[3]https://github.com/GermanMT/vexgen

exploits previos documentados. Para garantizar un entrenamiento efectivo, los datos se dividen en subconjuntos de entrenamiento, validación y prueba, asegurando que el modelo pueda generalizar sus predicciones a nuevos escenarios.

```
01: Esta es la vulnerabilidad "CVE-2021-44228".
02: Su descripción es: "Apache Log4j2 2.0-beta9 through 2.15.0 JNDI
features...".
03: El "part" es "a".
04: El "vendor" es "apache".
05: El "product" es"log4j".
06: El "version" es "2.0.1".
07: El "update" es "all".
08: El "edition" es "all".
09: El "language" es "all".
10: El "sw edition" es "all".
11: El "target sw" es "all".
12: El "target hw" es "all".
13: El "other" es "all".
14: El "attack vector" es "N".
15: El "attack complexity" es "L".
16: El "privileges required" es "N".
17: El "user interaction" es "N".
18: El "scope" es "C".
19: El "confidentiality" es "H".
20: El "integrity" es "H".
21: ¿Cuál es el mejor exploit? 50590
```

Figura 2. Un ejemplo de historia.

Para estructurar la información de una manera comprensible y uniforme, el sistema transforma los registros de vulnerabilidades en un formato de historias en lenguaje natural, siguiendo un esquema de pregunta-respuesta. Como se muestra en la Figura 2, cada historia sigue un patrón en el que la consulta es: "¿Cuál es el mejor exploit?", y la respuesta es el identificador del exploit más adecuado. Este enfoque permite que el modelo aprenda a identificar relaciones entre vulnerabilidades y exploits, mejorando su capacidad predictiva incluso en casos donde no existe un exploit registrado previamente.

El proceso de entrenamiento del modelo se basa en la creación de vocabularios específicos, la codificación de los datos y la retropropagación en una red neuronal optimizada para este tipo de tareas. Una vez entrenado, el modelo puede analizar nuevas vulnerabilidades y predecir el exploit más probable utilizando técnicas avanzadas de procesamiento del lenguaje natural y modelos neuronales.

La solución propuesta ha sido validada en múltiples conjuntos de datos de seguridad, aplicando validación cruzada para evaluar su precisión y capacidad de generalización. Los resultados muestran que el uso de SLMs en este contexto permite mejorar significativamente la identificación de vulnerabilidades explotables, facilitando la toma de decisiones en seguridad y optimizando la gestión de riesgos en la SSC.

III. Resultados Experimentales

En esta sección se muestran los resultados experimentales que validan la propuesta denominada Depex en la Subsección III-A, mientras que en las Subsecciones III-B y III-C, se describen los experimentos y las propuestas para su validación.

III-A. Resultados de Depex

Depex es una herramienta que automatiza la extracción de dependencias, la atribución de vulnerabilidades y la evaluación de riesgos mediante razonamiento automático basado en SMT. Gracias a su capacidad de analizar dependencias de múltiples lenguajes de programación (como Python, Java, JavaScript, Ruby o Rust), Depex es capaz de procesar aproximadamente 47.9 millones de repositorios públicos en GitHub, lo que representa el 11,4 % del total indexado en la plataforma.

III-A1. Comparación con Otras Herramientas: Para evaluar su desempeño, Depex fue comparado con Dependabot y Snyk en 93 proyectos de GitHub. Los resultados muestran que Depex supera significativamente a ambas herramientas en la detección de dependencias y vulnerabilidades. Frente a Dependabot, Depex encuentra un 64 % más dependencias y un 152 % más vulnerabilidades. En comparación con Snyk, la ventaja de Depex es aún mayor, detectando un 483 % más dependencias y un 204 % más vulnerabilidades. Además, Depex es la única herramienta que incorpora razonamiento automático, lo que permite a los desarrolladores comprender mejor la seguridad de sus ecosistemas de software.

III-A2. Evaluación Experimental: Para demostrar su impacto, se realizaron dos experimentos sobre 300 repositorios públicos, 100 proyectos por Python (PyPI), Java (Maven) y JavaScript (NPM). En el primer experimento, se midió la capacidad de Depex para generar grafos de dependencias en un máximo de 24 horas. Se logró completar los grafos para los proyectos en PyPI y NPM, mientras que en Maven, debido a la gran cantidad de dependencias y limitaciones en su API, no se alcanzó la profundidad máxima en ese tiempo, solo se pudo llegar a la capa 70 aproximadamente de dependencias transitivas.

En el segundo experimento, se analizó la eficiencia del razonamiento automático. Se determinó que 98 % de los proyectos fueron evaluados en tiempos razonables, y aproximadamente 15 % de los repositorios tenían al menos una configuración vulnerable. De los proyectos libres de vulnerabilidades, 81 % se mantuvieron sin configuraciones inseguras, mientras que el 19 % tenían al menos una configuración vulnerable. Esto demuestra que Depex no solo identifica dependencias y vulnerabilidades con mayor precisión, sino que también proporciona un análisis avanzado para la mitigación de riesgos.

En la Tabla I se muestra un ejemplo concreto, el impacto de las configuraciones de versiones en *RapidSMS* a lo largo del proceso de análisis. Los resultados mostraron que en su configuración mínima, *RapidSMS* no presentaba vulnerabilidades, con un impacto registrado de 0.0. Sin embargo, en su configuración máxima, el impacto alcanzó 3.99, indicando que ciertas dependencias podían generar configuraciones inseguras. En cuanto al tiempo de ejecución, se observó una estabilidad notable, con un tiempo de procesamiento mínimo de 0.28 segundos y un tiempo máximo de 0.27 segundos, lo que sugiere que el rendimiento del análisis se mantiene estable en diferentes escenarios.

III-B. Resultados de VEXGen

El diseño de este estudio experimental tiene como objetivo evaluar la capacidad de la propuesta para generar archivos VEX y eVEX de manera automatizada, facilitando la identificación y gestión de vulnerabilidades en software de terceros en una SSC. Se pretende analizar la cantidad de información generada, su eficiencia en términos de tiempo de procesamiento y la utilidad de agrupar y priorizar declaraciones para mejorar la toma de decisiones en seguridad.

Tabla I
EJEMPLO CONCRETO DEL PROYECTO *RapidSMS*.

Nombre	Impacto Mínimo	Tiempo Mínimo (s)	Impacto Máximo	Tiempo Máximo (s)	N.º Dependencias	N.º Vulnerabilidades	N.º Exploits
RapidSMS	0.0	0.28	3.99	0.27	54	330	38

Para ello, se emplearán 50 SBOM de proyectos reales alojados en GitHub, asegurando que los casos de prueba reflejen escenarios prácticos. La evaluación incluirá la capacidad del sistema para documentar vulnerabilidades desde la NVD y OSV, y proporcionar información complementaria en archivos eVEX cuando no se pueda determinar automáticamente el estado de una vulnerabilidad. La información complementaria se extraerá de VulnersDB para exploits, y de la página del MITRE para los CWE. También se analizará el impacto del número de dependencias en el tiempo de procesamiento y se estudiará la viabilidad del sistema en entornos con gran volumen de datos.

Además, se explorará cómo el agrupamiento y priorización de declaraciones puede mejorar la eficiencia en la gestión de riesgos. Se investigará la clasificación de vulnerabilidades según CWE y la priorización basada en factores como el vector de ataque y los requisitos de autenticación, con el fin de facilitar la identificación de amenazas críticas.

Estos experimentos permitirán validar la capacidad de VEXGen para mejorar la automatización en la detección y gestión de vulnerabilidades, optimizando la trazabilidad y la mitigación de riesgos en la SSC.

III-C. Resultados del modelo SLM

Para evaluar la capacidad de nuestro modelo SLM en la predicción de exploits para vulnerabilidades, se diseñará un estudio experimental basado en el uso de modelos de lenguaje pequeño (SLM). La experimentación tendrá como objetivo analizar el desempeño del modelo en distintos escenarios, considerando su eficiencia, precisión y estabilidad en entornos con variaciones lingüísticas.

Se plantea utilizar datos reales obtenidos de bases de datos de seguridad como NVD y ExploitDB, con los cuales se entrenará y validará el modelo. Actualmente, se dispone de 300.000 vulnerabilidades de NVD y 50.000 exploits de ExploitDB aproximadamente. Además, según la NVD, del total de vulnerabilidades aproximadamente sólo el 9 % tienen exploits conocidos en ExploitDB. Por otro lado, aproximadamente el 63 % de los exploits de ExploitDB tienen un CVE asociado. A través de estos experimentos, se busca determinar la precisión del modelo SLM propuesto para predecir correctamente la existencia de exploits para vulnerabilidades documentadas y cómo su rendimiento varía en función de la cantidad de datos utilizados en el entrenamiento.

III-C1. Diseño Experimental: El modelo será implementado en Python 3.10, utilizando bibliotecas especializadas en aprendizaje profundo como TensorFlow, Keras y NumPy. Para su ejecución, se prevé el uso de sistemas con capacidades computacionales avanzadas, incluyendo procesadores de alto rendimiento y tarjetas gráficas optimizadas para entrenamiento de modelos neuronales. El estudio se llevará a cabo en dos escenarios experimentales:

- **Escenario A:** Se entrenará el modelo con un conjunto de datos más grande, asignando el 50 % de los datos para entrenamiento, el 25 % para validación y prueba respectivamente.
- **Escenario B:** Se reducirá significativamente la cantidad de datos de entrenamiento, utilizando solo un 10 % para entrenamiento y validación, mientras que el 80 % se reservará para prueba.

Este diseño permitirá analizar cómo afecta la reducción de datos al rendimiento del modelo, tanto en términos de precisión como de tiempos de entrenamiento e inferencia. Además, se explorará si el modelo es capaz de ajustarse a volúmenes de datos más pequeños sin comprometer su capacidad predictiva.

III-C2. Métricas de Evaluación: Para evaluar el desempeño de nuestro modelo SLM, se analizarán tres aspectos clave:

1. **Eficiencia:** Se medirá el tiempo requerido para entrenar el modelo en cada escenario, así como el tiempo de inferencia en la fase de prueba. Se espera determinar si la reducción del conjunto de entrenamiento permite una optimización del tiempo de procesamiento sin afectar significativamente la calidad de las predicciones.
2. **Precisión:** Se calcularán métricas de evaluación como la exactitud y el coeficiente de correlación de Matthew (MCC) para medir la capacidad del modelo en la predicción de exploits. Se analizará si una menor cantidad de datos de entrenamiento impacta de manera significativa la precisión de las predicciones.
3. **Impacto del Idioma:** Se examinará el rendimiento del modelo en diferentes idiomas (inglés, español, alemán, hindi y ruso) para determinar si la variabilidad lingüística afecta su capacidad de predicción.

Además, se llevará a cabo un análisis estadístico para validar la estabilidad de los resultados, verificando la distribución de las métricas de rendimiento mediante pruebas de normalidad y homocedasticidad. También se compararán los tiempos de procesamiento y la precisión entre los dos escenarios experimentales para identificar diferencias significativas.

IV. TRABAJOS RELACIONADOS

El análisis de dependencias en la cadena de suministro de software ha sido ampliamente estudiado, tanto para dependencias directas [17] como indirectas [18]. Se han desarrollado diversas estrategias para evaluar el impacto de estas dependencias en la seguridad del software, incluyendo el uso de grafos de dependencias para representar relaciones y evaluar riesgos [19]. Algunos enfoques identifican vulnerabilidades en dependencias comparando nombres y términos con bases de datos de CVEs y CPEs [20], [21], mientras que otros aplican análisis semántico [22], inspección de código [23], [24] y aprendizaje profundo [25] para mejorar la detección de riesgos.

El Software Bill of Materials (SBOM) ha surgido como una herramienta clave para la gestión de dependencias y la seguridad en la cadena de suministro de software. Estudios

recientes han analizado su precisión y la efectividad de las herramientas que los generan, identificando problemas en la detección de componentes de terceros y la evaluación de vulnerabilidades [26], [27], [28]. Además, se ha señalado que la generación de VEX (Vulnerability Exploitability eXchange) sigue siendo un desafío debido a su naturaleza manual y propensa a errores [29], [30]. Algunas herramientas comerciales, como vexctl y vexy, facilitan la creación de archivos VEX, pero sin automatizar el análisis de explotabilidad [31], [32]. La única herramienta con generación automática de VEX, Kubescape, está restringida a entornos Kubernetes [33].

Desde la perspectiva del aprendizaje automático (ML), varios estudios han explorado la predicción de explotabilidad de vulnerabilidades a partir de la información de CVEs [34], [35], [36]. Algunos enfoques aplican modelos de clasificación para predecir la existencia de exploits [37], [38], [39], [40], mientras que otros utilizan técnicas de NLP y modelos de lenguaje de gran escala (LLMs) para generar descripciones y pruebas de explotabilidad [41], [42]. Herramientas como PENTESTGPT [43] y Sec-PaLM2 [44] han sido diseñadas para asistir en tareas de ciberseguridad, desde pruebas de penetración hasta generación de parches de seguridad [45], [46], [47].

Finalmente, en el ámbito de la Generación Automática de Exploits (AEG), los enfoques tradicionales han empleado ejecución simbólica para descubrir rutas explotables en el código [48]. Investigaciones recientes han intentado mejorar estos métodos optimizando la búsqueda de vulnerabilidades explotables mediante algoritmos avanzados [49], [50], aunque aún no existe una solución completamente automatizada que integre la predicción de explotabilidad, el análisis de dependencias y la generación de exploits en un mismo enfoque.

V. Conclusiones y Trabajos Futuros

La seguridad de la SSC es un desafío crítico en la protección de ecosistemas digitales, especialmente debido al incremento en ataques dirigidos a dependencias y la creciente complejidad en la gestión de vulnerabilidades. En este trabajo, hemos propuesto un conjunto de soluciones que abordan este problema que afecta a diferentes partes de la SSC.

El análisis de SBOM y el modelado de dependencias proporcionan a los equipos de seguridad obtener una visión más clara de la estructura de la SSC, lo que les permite detectar de forma temprana temprana de vulnerabilidades en dependencias directas e indirectas. La automatización en la generación de archivos VEX y eVEX reduce considerablemente el esfuerzo manual en la gestión de vulnerabilidades, proporcionando información detallada sobre su explotabilidad en contextos específicos y mejorando la trazabilidad de los riesgos en proyectos de software. Además, la implementación de modelos de aprendizaje automático para la predicción y recomendación de exploits facilita la priorización de amenazas con mayor potencial de explotación, optimizando la respuesta ante posibles incidentes de seguridad y agilizando la toma de decisiones.

No obstante, es importante destacar que las herramientas desarrolladas también presentan ciertas limitaciones. Su adopción en entornos empresariales reales puede verse condicionada por factores como la disponibilidad de recursos computacionales, la integración con infraestructuras preexistentes o la necesidad de mantener actualizadas las bases de datos de vulnerabilidades. Además, el razonamiento automático basado en SMT puede conllevar tiempos de procesamiento elevados en proyectos con estructuras de dependencias muy complejas. Estos aspectos deben ser tenidos en cuenta al aplicar las soluciones propuestas en contextos operativos con restricciones técnicas o presupuestarias.

El impacto de estas propuestas se reflejará en una gestión de riesgos más eficiente, reduciendo los tiempos de detección y mitigación de vulnerabilidades dentro de la SSC. La integración de estos enfoques permitirá a las organizaciones fortalecer su postura de seguridad, minimizando la exposición a ataques y mejorando la resiliencia de sus ecosistemas software. Como trabajo futuro, exploraremos nuevas técnicas de modelado y análisis de seguridad que permitan reforzar aún más la ciberseguridad en la SSC, con el objetivo de ofrecer soluciones escalables y adaptables a diferentes entornos de desarrollo y despliegue.

Agradecimientos

Este trabajo ha sido financiado por los proyectos KOSMOS-US PID2024-155363OB-C42, y AETHER-US PID2020-112540RB-C44, y ALBA-US TED2021-130355B-C32 financiadas por el MICIU/AEI/10.13039/501100011033 y la «NextGenerationEU/PRTR» de la Unión Europea.

Referencias

[1] T. Zimmermann, C. Bird, and N. Nagappan, "The software supply chain: The source of security issues?" *IEEE Software*, vol. 39, no. 1, pp. 16–22, 2022.

[2] R. He, H. He, Y. Zhang, and M. Zhou, "Automating dependency updates in practice: An exploratory study on github dependabot," *IEEE Transactions on Software Engineering*, vol. 49, no. 8, pp. 4004–4022, 2023.

[3] W. Dormann, "Understanding and managing software bill of materials (sbom)," *CISA Cybersecurity Report*, 2021. [Online]. Available: https://www.cisa.gov/publication/software-bill-materials

[4] P. Foreman, *Vulnerability management*. Auerbach Publications, 2019.

[5] P. Engebretson, *The basics of hacking and penetration testing: ethical hacking and penetration testing made easy*. Elsevier, 2013.

[6] M. Ohm, T. Fiebig, K. Krombholz, and T. Holz, "Backstabber's knife collection: A review of open source software supply chain attacks," in *Proceedings of the USENIX Security Symposium*, 2020, pp. 1965–1982.

[7] M. Brennan, J. Spring, and A. Householder, "Supply chain attacks: Understanding the threats," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–36, 2022.

[8] J. A. Galindo, D. Benavides, and S. Segura, "Debian packages repositories as software product line models. towards automated analysis," in *ACOTA, Belgium, September, 2010*, vol. 688. CEUR-WS.org, 2010, pp. 29–34.

[9] R. Cox, "Surviving software dependencies," *Commun. ACM*, vol. 62, no. 9, p. 36–43, 8 2019. [Online]. Available: https://doi.org/10.1145/3347446

[10] P. J. Morrison, R. Pandita, X. Xiao, R. Chillarege, and L. Williams, "Are vulnerabilities discovered and resolved like other defects?" *Empirical Software Engineering*, vol. 23, no. 3, pp. 1383–1421, 6 2018. [Online]. Available: https://doi.org/10.1007/s10664-017-9541-1

[11] S. Dass and A. S. Namin, "Vulnerability coverage for adequacy security testing," in *35th Annual ACM Symposium on Applied Computing*, ser. SAC '20. New York, NY, USA: ACM, 2020, p. 540–543. [Online]. Available: https://doi.org/10.1145/3341105.3374099

[12] S. Elder, M. R. Rahman, G. Fringer, K. Kapoor, and L. Williams, "A survey on software vulnerability exploitability assessment," *ACM Comput. Surv.*, vol. 56, no. 8, Apr. 2024. [Online]. Available: https://doi.org/10.1145/3648610

[13] Cybersecurity and Infrastructure Security Agency (CISA), "Vulnerability Exploitability eXchange (VEX) – Use Cases," U.S. Department of Homeland Security, Tech. Rep., Apr. 2022, accessed: 2025-03-20. [Online]. Available: https://www.cisa.gov/sites/default/files/publications/VEX_Use_Cases_April2022.pdf

[14] R. Hiesgen, M. Nawrocki, T. C. Schmidt, and M. Wählisch, "The race to the vulnerable: Measuring the log4j shell incident," 2022.

[15] "log4shell," https://github.com/NCSC-NL/log4shell/blob/main/software/, 2023.

[16] A. Germán Márquez, Ángel Jesús Varela-Vaca, M. T. Gómez López, J. A. Galindo, and D. Benavides, "Vulnerability impact analysis in software project dependencies based on satisfiability modulo theories (smt)," *Computers & Security*, vol. 139, p. 103669, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404823005795

[17] J. Cox, E. Bouwers, M. van Eekelen, and J. Visser, "Measuring dependency freshness in software systems," in *2015 IEEE/ACM 37th ICSE*, vol. 2, 2015, pp. 109–118.

[18] R. G. Kula, D. M. German, A. Ouni, T. Ishio, and K. Inoue, "Do developers update their library dependencies?" *Empirical Softw. Engg.*, vol. 23, no. 1, p. 384–417, Feb. 2018. [Online]. Available: https://doi.org/10.1007/s10664-017-9521-5

[19] Q. Li, J. Song, D. Tan, H. Wang, and J. Liu, "Pdgraph: A large-scale empirical study on project dependency of security vulnerabilities," in *Proceedings of the 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2021)*, 2021, Conference paper, p. 161 – 173, cited by: 3. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114898505&doi=10.1109DSN48987.2021.00031&partnerID=40&md5=84dd1aa636a487a5ea08a2e4328433f5

[20] M. Cadariu, E. Bouwers, J. Visser, and A. van Deursen, "Tracking known security vulnerabilities in proprietary software systems," in *2015 IEEE 22nd SANER*, 2015, pp. 516–519.

[21] J. Hejderup, "In dependencies we trust: How vulnerable are dependencies in software modules?" in *Proceedings of Delft University of Technology Research*, 2015.

[22] S. S. Alqahtani, E. E. Eghan, and J. Rilling, "Tracing known security vulnerabilities in software repositories – a semantic web enabled modeling approach," *Science of Computer Programming*, vol. 121, pp. 153–175, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167642316000253

[23] S. Ponta, H. Plate, and A. Sabetta, "Beyond metadata: Code-centric and usage-based analysis of known vulnerabilities in open-source software," in *2018 ICSME*. Los Alamitos, CA, USA: IEEE Computer Society, sep 2018, pp. 449–460. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICSME.2018.00054

[24] R. Ye, L. Liu, S. Hu, F. Zhu, J. Yang, and F. Wang, "Jslim: Reducing the known vulnerabilities of javascript application by debloating," *Communications in Computer and Information Science*, vol. 1403 CCIS, p. 128 – 143, 2022, cited by: 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124125136&doi=10.1007978-3-030-93956-4_8&partnerID=40&md5=af60b72693fa86ed6d588b29cd01b380

[25] X. Yuan, G. Lin, Y. Tai, and J. Zhang, "Deep neural embedding for software vulnerability discovery: Comparison and optimization," *Security and Communication Networks*, vol. 2022, 2022, cited by: 1; All Open Access, Gold Open Access. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124145032&doi=10.115520225203217&partnerID=40&md5=d85c515814070b05ff84f3d24d82ff18

[26] T. Bi, B. Xia, Z. Xing, Q. Lu, and L. Zhu, "On the way to sboms: Investigating design issues and solutions in practice," *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 6, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3654442

[27] M. Mirakhorli, D. Garcia, S. Dillon, K. Laporte, M. Morrison, H. Lu, V. Koscinski, and C. Enoch, "A landscape study of open source and proprietary tools for software bill of materials (sbom)," 2024. [Online]. Available: https://arxiv.org/abs/2402.11151

[28] G. Benedetti, S. Cofano, A. Brighente, and M. Conti, "The impact of sbom generators on vulnerability assessment in python: A comparison and a novel approach," 2024. [Online]. Available: https://arxiv.org/abs/2409.06390

[29] N. Zahan, E. Lin, M. Tamanna, W. Enck, and L. Williams, "Software bills of materials are required. are we there yet?" *IEEE Security & Privacy*, vol. 21, no. 2, pp. 82–88, 2023.

[30] B. Xia, T. Bi, Z. Xing, Q. Lu, and L. Zhu, "An empirical study on software bill of materials: Where we stand and the road ahead," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2023, pp. 2630–2642.

[31] "Vexctl," https://github.com/openvex/vexctl, 2024.

[32] "Vexy," https://github.com/madpah/vexy, 2024.

[33] "Linux foundation," https://kubescape.io/blog/2023/12/07/kubescape-support-for-vex-generation/#enter-kubescape, 2024.

[34] N. Tavabi, P. Goyal, M. Almukaynizi, P. Shakarian, and K. Lerman, "Darkembed: Exploit prediction with neural language models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[35] M. Aota, H. Kanehara, M. Kubo, N. Murata, B. Sun, and T. Takahashi, "Automation of vulnerability classification from its description using machine learning," in *2020 IEEE Symposium on Computers and Communications (ISCC)*, 2020, pp. 1–7.

[36] M. Edkrantz, "Predicting exploit likelihood for cyber vulnerabilities with machine learning," in *Computer Science*, 2015.

[37] E. Iannone, G. Sellitto, E. Iaccarino, F. Ferrucci, A. De Lucia, and F. Palomba, "Early and realistic exploitability prediction of just-disclosed software vulnerabilities: How reliable can it be?" *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 6, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3654443

[38] H. Eskandari, M. Bewong, M. Geaur Rahman, and S. Ur Rehman, "Outcentr: A method for predicting exploits of cyber vulnerabilities in high dimensional datasets," *IEEE Access*, vol. 12, pp. 133 030–133 044, 2024.

[39] C. Figueiredo Pereira, J. Gabriel Lopes de Oliveira, R. Azevedo Santos, D. Vieira, L. Miranda, G. Zaverucha, L. Pfleger de Aguiar, and D. Sadoc Menasché, "A statistical relational learning approach towards products, software vulnerabilities and exploits," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 3782–3802, 2023.

[40] K. Charmanas, N. Mittas, and L. Angelis, "Exploitation of vulnerabilities: A topic-based machine learning framework for explaining and predicting exploitation," *Information*, vol. 14, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/7/403

[41] G. Yang, Y. Zhou, X. Chen, X. Zhang, T. Han, and T. Chen, "Exploitgen: Template-augmented exploit code generation based on codebert," *Journal of Systems and Software*, vol. 197, p. 111577, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121222002539

[42] X. Ruan, Y. Yu, W. Ma, and B. Cai, "Prompt learning for developing software exploits," in *Proceedings of the 14th Asia-Pacific Symposium on Internetware*, ser. Internetware '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 154–164. [Online]. Available: https://doi.org/10.1145/3609437.3609445

[43] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass, "Pentestgpt: An llm-empowered automatic penetration testing tool," 2023.

[44] "Google PaLM 2," https://ai.google/discover/palm2/, 2024.

[45] D. Noever, "Can large language models find and fix vulnerable software?" 2023.

[46] N. T. Islam, J. Khoury, A. Seong, G. D. L. T. Parra, E. Bou-Harb, and P. Najafirad, "Llm-powered code vulnerability repair with reinforcement learning and semantic reward," 2024.

[47] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining zero-shot vulnerability repair with large language models," 2022.

[48] S. K. Cha, T. Avgerinos, A. Rebert, and D. Brumley, "Unleashing mayhem on binary code," in *2012 IEEE Symposium on Security and Privacy*, 2012, pp. 380–394.

[49] R. Wang, K. Chen, Z. Pan, Y. Li, Q. Li, Y. Li, M. Zhang, and C. Zhang, "Tunter: Assessing exploitability of vulnerabilities with taint-guided exploitable states exploration," *Computers & Security*, vol. 124, p. 102995, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740482200387X

[50] J. Liu, H. An, J. Li, and H. Liang, "Detecting exploit primitives automatically for heap vulnerabilities on binary programs," 2022.

Simulating Cyberattacks through a Breach Attack Simulation (BAS) Platform empowered by Security Chaos Engineering (SCE)

Arturo Sánchez-Matas[1], Pablo Escribano Ruiz[1], Daniel Díaz-López[1,2]
Ángel Luis Perales Gómez[3], Pantaleone Nespoli[1], Gregorio Martínez Pérez[1]
[1]Department of Information and Communications Engineering, University of Murcia, 30100, Murcia, Spain
{arturo.sanchezm, pablo.escribanor, danielorlando.diaz, pantaleone.nespoli, gregorio}@um.es
[2]School of Engineering, Science and Technology, Universidad del Rosario, Bogotá, Colombia
danielo.diaz@urosario.edu.co
[3]Department of Computers Engineering and Technology, University of Murcia, 30100, Murcia, Spain
angelluis.perales@um.es

***Abstract*—In today's digital landscape, organizations face constantly evolving cyber threats, making it essential to discover slippery attack vectors through novel techniques like Security Chaos Engineering (SCE), which allows teams to test defenses and identify vulnerabilities effectively. This paper proposes to integrate SCE into Breach Attack Simulation (BAS) platforms, leveraging adversary profiles and abilities from existing threat intelligence databases. This innovative proposal for cyberattack simulation employs a structured architecture composed of three layers: SCE Orchestrator, Connector, and BAS layers. Utilizing MITRE Caldera in the BAS layer, our proposal executes automated attack sequences, creating inferred attack trees from adversary profiles. Our proposal's evaluation illustrates how integrating SCE with BAS can enhance the effectiveness of attack simulations beyond traditional scenarios, and be a useful component of a cyber defense strategy.**

***Index Terms*—Breach Attack Simulation, adversary profiling, attack trees, Security Chaos Engineering, threat intelligence, Cyber Situational Awareness, cyber defense**

Contribution type: *Original research*

I. Introduction

Cyber defense encompasses the strategies, technologies, and processes designed to protect networks, devices, programs, and data from cyberattacks [1]. It involves implementing security measures such as firewalls, intrusion detection systems, and encryption to safeguard information integrity, confidentiality, and availability. Simulating cyberattacks is crucial in cyber defense as it enables organizations to proactively identify and address vulnerabilities before they can be exploited by adversaries [2]. Through realistic attack simulations, defenders can evaluate the effectiveness of their security controls, improve incident response strategies, and enhance overall resilience against evolving cyber threats. This proactive approach ensures that defense mechanisms are robust and adaptable, ultimately strengthening an organization's security posture.

Current methods for simulating adversary behavior in cyberspace primarily involve the utilization of automated frameworks and deception techniques to replicate known attack strategies [3]. Tools such as adversary emulation frameworks are specifically designed to mimic the actions of potential attackers and assess the effectiveness of existing security measures [4]. Additionally, Honeypots and Threat Intelligence Platforms are crucial in profiling adversary behavior by attracting attackers and providing detailed information about their Tactics, Techniques, and Procedures (TTPs). While Honeypots capture and analyze the actions of real adversaries, Threat Intelligence Platforms aggregate indicators of compromise and attack vectors from current and past campaigns, aiding in the development of accurate adversary profiles. However, traditional simulation methods face significant challenges, including the difficulty of adapting to rapidly evolving threat landscapes and accurately replicating the sophisticated tactics of advanced adversaries. These limitations can result in incomplete assessments of an organization's security posture, leaving critical vulnerabilities undiscovered and unaddressed.

In this scenario, Breach Attack Simulation (BAS) platforms are essential for enhancing cyber defense strategies by enabling organizations to simulate realistic cyberattacks in a controlled environment [5]. That is, BAS facilitates the creation of comprehensive and repeatable attack scenarios that help security teams evaluate the effectiveness of their defenses, identify vulnerabilities, and improve incident response protocols. Various BAS solutions are available in the market, including proprietary and open-source platforms. Proprietary BAS tools, such as SafeBreach and AttackIQ, offer extensive features and dedicated support, providing organizations with robust capabilities to simulate complex attack vectors. Open-source frameworks such as MITRE Caldera and Atomic Red Team provide adaptability and extensibility, allowing security teams to tailor simulations to their specific needs [6].

A novel methodology, known as Security Chaos Engineering (SCE), used to evaluate security in systems emerged in the last years and is winning popularity in industry and academy. SCE is a transformative approach that applies the principles

of chaos engineering to the security domain. It involves deliberately injecting faults or simulated attacks into a live system to observe how it behaves under adverse conditions. By intentionally triggering failures, SCE aims to expose hidden vulnerabilities and validate the resilience of complex, modern software systems. SCE experiments can range from simulating network congestion to testing the limits of authentication mechanisms, with each test designed to mimic real-world attack scenarios. As a result, SCE transforms a reactive security stance into a proactive and adaptive one, fostering a culture of continuous learning and improvement [7].

The main objective of our research is to enhance organizational defenses by proactively identifying vulnerabilities, thereby enabling continuous improvement in their security posture. To achieve this, we design an innovative framework that utilizes a BAS platform and SCE as integrated means. In this paper, we present the design of our framework, demonstrating how these integrated means facilitate a comprehensive approach to proactive security testing. All of the material developed in our proposal is available as open-source code in the project repository.

The main contributions of this paper are summarized as follows:

- Proposal of a novel method to simulate adversary behavior through the integration of a BAS platform with a SCE methodology. This seamless combination allows for dynamic and realistic security assessments by leveraging the strengths of both BAS platforms and SCE principles.
- The composition of Attack Trees (AT) that leverage existing databases of TTPs to facilitate realistic simulations. By utilizing comprehensive TTP databases, our framework constructs detailed and accurate attack scenarios that mirror real-world adversary strategies.
- The validation of our proposal in a simulated scenario composed by a critical-target information system attacked by an advanced adversary, who is known for having access to a TTP database. This simulation demonstrates that our proposal is reliable and effective in identifying the possible attack vectors in a target systems.

This paper is structured as follows: Section II reviews the most recent works related to BAS and SCE, evaluating their strengths and weaknesses. Next, Section III introduces our proposed framework for simulating cyberattacks. In Section IV, we describe the experiments performed using our framework to validate its effectiveness. Finally, Section V concludes the paper, discussing the implications of our findings and outlining directions for future research.

II. State of the art

In this section we will present several research papers related to BAS, attack trees and SCE, which are relevant to understand the novelty of our proposal.

In particular, an attack emulation proposal, named SpecRec, applicable to a variety of steps of a cyberattack was described in [8]. It analyzes white papers related to known groups or Advanced Persistent Threats (APT), like Mirai, to extract attack specification that will be emulated later. SpecRep leverages an LLM to translate selected white papers into high-level attack specifications, which are used internally to define the attack's logic. SpecRec also defines a metalanguage to extract attack objectives and a compiler to build attack scenarios. In this way, SpecRep is analog to a BAS platform, where it can generate new attack flows, and is not limited to a predefined playbook of attacks.

Kijong Koo et al. [9] proposed a novel machine learning-based method to generate attack graphs to predict attack paths. In order to perform this task, two stages were set. The first one revolves around learning the attack paths from known databases relying on machine learning, and the second one consists of generating attack graphs using information such as the network topology and the systems in it. This approach promises to tackle the main issue around the generation of attack graphs, which is the amount of time required to build them manually.

Basit Ajmal et al. [10] proposed an approach for adversary emulation based on the MITRE ATT&CK framework to provide an effective way of testing the defense of cyber systems, allowing the profiling of adversaries. Such a paper explored the idea of automated attack generation, facilitating repeated tests and mimicking the behavior of a real-world attack. The simulation developed in such proposal is performed through several means, depending on the attack to be performed, using scripts and adapting payloads.

In a similar way as the previous related work, Sang Ho Oh et al. [11] proposed the use of a Deep Reinforcement Learning (DRL) algorithm to simulate dynamic cyberattacks in scenarios derived from the MITRE ATT&CK framework. This algorithm is applied to agents that simulate cyberattacks in an environment that mimics a real-world scenario with real vulnerabilities. The DRL agents are trained to adapt on the fly and change their strategies depending on the interactions within the scenario, improving the accuracy of the simulations.

On the other side, we have ChaosXploit, which is a SCE framework proposed by Palacios et al. [12], that facilitates the implementation of the SCE methodology through different experiments. ChaosXploit leverages attack trees to perform experiments attacking cloud infrastructure, such as AWS S3 buckets. This is done thanks to a "hypothesis generator", which forms part of a "knowledge database", an "observer" who has knowledge about the system and the experiments being run, and an "experiment runner", in charge of running the experiments.

Another application of SCE was found in [13]. Such a paper studies the security of the traditional DevSecOps lifecycle, arguing that new software vulnerabilities might appear as early as the design phases. It complements the conventional DevSecOps practice through the use of LLMs to automate threat discovery. This proposal also applies the SCE methodology to execute more complex and chaotic tests, which would not be possible with more conventional or standard approaches.

After analyzing the previous papers, we realized that there was no previous proposal combining both BAS and SCE. BAS platforms suffer from a lack of variety when it comes to choosing the attack vectors to use, and a failure on a chosen vector will often lead to the attack ending. Thus, integrating BAS and SCE sounds like a promising idea that would allow to compose attack trees from the information contained in a threat intelligence database. In addition, cyberattacks would benefit from the complexity and freedom of SCE, without suffering the rigidity of a BAS platform that is limited to a set of attack vectors.

III. Proposal of Attack Simulations empowered by Security Chaos Engineering

This section presents our proposed solution for integrating SCE into adversary simulation workflows. The architecture of the proposal, illustrated in Figure 1, is organized into three interconnected layers: i) *BAS layer*, which is based on MITRE Caldera and executes automated attack sequences using real-world adversary profiles, ii) *Connector layer*, which is a middleware layer that interfaces with external platforms and standardizes communication; and iii) *SCE Orchestrator layer*, who is responsible for designing and managing chaos-driven attack scenarios. Together, these 3 layers facilitate systematic and adaptive testing of defenses by combining the unpredictability of SCE with the precision of BAS tools. The following subsections detail the design, functionality, and integration of each component within this modular framework.

A. BAS layer

The BAS layer allows us to run controlled and defined attacks on targets of our choice, giving the possibility of checking what the systems are vulnerable to, and seeing how they were attacked. Our *BAS layer* is implemented with MITRE Caldera, which is an open source BAS solution developed by MITRE. The Caldera module behind the decision-making process was adjusted to allow the SCE methodology to work in conjunction with it, leveraging some native Caldera modules, such as agents, operations, and abilities, which are vital for the proposal and provide the groundwork from which we have expanded.

1) Core System: The core system, which lays at the foundation of Caldera, allows the management of each of its components, like campaigns, operations, adversaries, and abilities, as well as the intercommunication between each of them. This core system offers an API connection so the *connector layer* can communicate with it via API requests.

2) Agents: To provide proper attacker simulation, Caldera offers a selection of different agents, i.e. attackers, which are not so different in functionalities from a Remote Access Trojan (RAT). These agents communicate with the Caldera core system, which acts as a Command and Control (C2) server, providing information about the target, reporting result of the actions they perform, and receiving orders about the attacks to perform. Different parameters of the agents can be adjusted to simulate real attackers, like sleep timers, which are useful when trying to avoid detection.

3) Abilities: (referred to as Technique in MITRE TTP Matrix) An ability is a term used to describe the action that an agent is going to take, the instruction it is going to receive, and even the attack that it is going to perform; all of the terms can be used interchangeably.

4) Adversary profiles: An adversary profile determines which actions will be taken in an attack from start to finish to achieve an attack goal, which could be data ex-filtration, discovery, or many other options. Thus, an adversary profile is comprised of several abilities, and a user can choose to run the abilities he/she deems necessary.

5) Operations: An operation is a "scene" of sorts, a playground for Caldera to work in. In our proposal, we use a blank operation which does not contain any adversary profile so that the *SCE Orchestrator layer* can decide which actions to execute freely.

6) Campaigns: Caldera offers us the possibility of managing a scenario with offensive and defensive actions simultaneously. This scenario is referred to as a campaign, where multiple operations, ones for defense and ones for offense are executed by red and blue agents. Each campaign can mirror real-life scenarios, allowing to test the resilience of an organization via a controlled and repeatable exposure to multiple campaigns.

7) Plugins: Finally, Caldera allows for easy extensibility through plugins which are software components addressed to do specific tasks, e.g. allowing to view the interaction between blue and red agents, simulating human behavior in a target machine, among others.

B. Connector layer

The *Connector layer* is responsible for ensuring communication between the *SCE Orchestrator layer* and the *BAS layer*. Whenever a specific branch of the attack tree is chosen by the *SCE orchestrator layer*, the *connector layer* will execute the abilities associated with this branch, for which it will use a set of API requests addressed to different endpoints available in the *BAS layer*. The *connector layer* is composed of a set of modules described next.

1) State Fetcher: The state fetcher enables the initial information recollection step of our proposal, allowing us to obtain the general state of the BAS before progressing further. This module permits a swift collection of relevant *BAS layer* information, such as the agents deployed, available abilities, and existing adversary profiles. The collected information mentioned is used to create an "operation", which allows the execution of a selected "adversary profile" ant its "abilities", through an agent deployed in the target.

2) Operation creator: All actions in the BAS have to exist in the context of an operation, so this module ensures the creation of a blank operation in which all subsequent decisions can be made, like a blank canvas waiting for a painter.

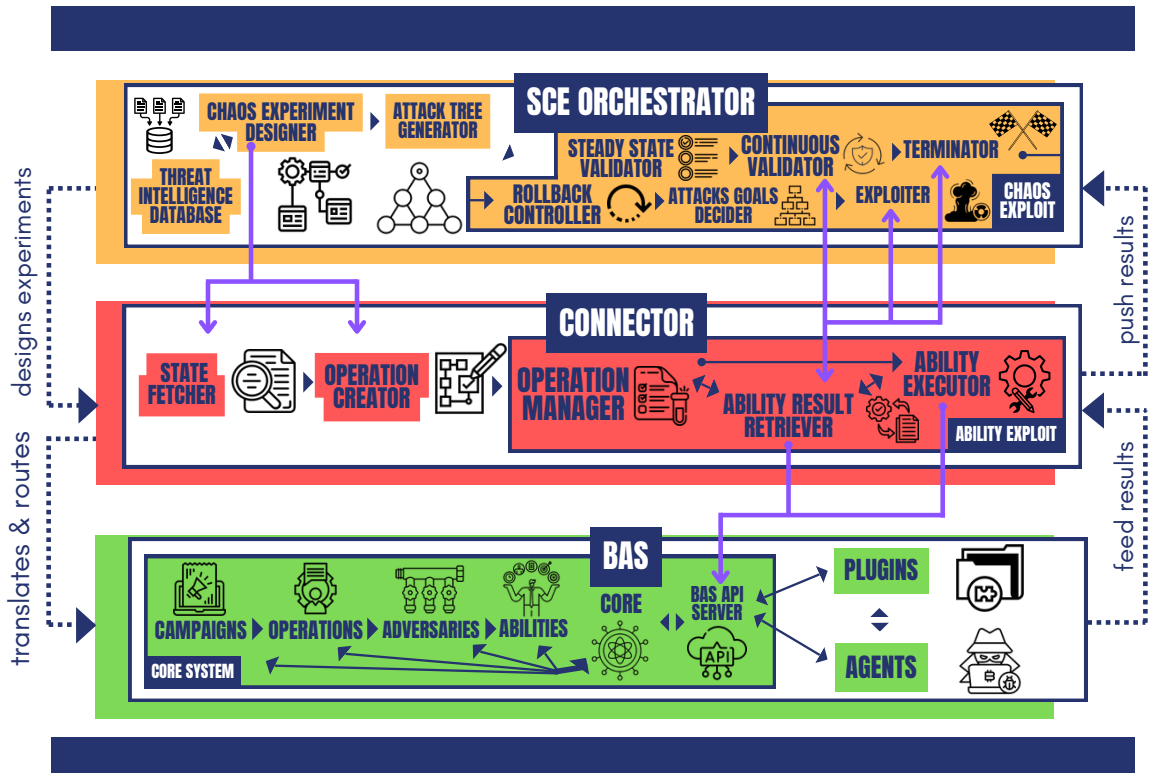


Figure 1: Three-layered architecture combining BAS with SCE

3) Operation Manager: Once the operation is created, there is a certain degree of settings that can be modified before running the experiments. These settings can range from a set of actions, such as modifying the sleep timer of agents, the trusted timer to mark agents as untrustworthy, or, more importantly for the experiments, mutating the operation adding new abilities to execute. These abilities are called potential links, which is a technical term for an ability or technique, and is a way of telling the operation what abilities should be executed.

4) Ability executor: After having a blank operation, we can proceed by traversing the attack tree and selecting the next action. For these actions to be performed, the ability executor is in charge of performing the communication with the *BAS layer*'s API to request the execution of an action, i.e., an "ability".

5) Ability result retriever: Once an ability is executed, the result is not available immediately. Thus, the ability result retriever "polls" the *BAS layer* to validate if a result has been generated, facilitating the recovery of the output of an attack. The reason for this behavior is that the agents do not perform actions in real time, but according to timers that simulate real threat agents that might want to remain inconspicuous.

C. SCE Orchestrator layer

The *SCE Orchestrator layer* is the one devoted for planning and managing chaos-based attack simulations. In particular, it is responsible for accepting user inputs (target machine (t_i), agent to use (a_i), number of parallel experiments (n), and security objective (o_i)) to create customized attack scenarios. Leveraging these inputs, it oversees the execution of experiments, ensuring efficient alignment with overall security strategies. This layer manages the entire simulation lifecycle, from setup to teardown, while maintaining consistency across multiple runs. Additionally, it incorporates safety mechanisms to monitor system stability during tests and automatically reverts actions if unexpected issues occur. By transforming security goals into structured chaos experiments, the orchestrator enables teams to evaluate and enhance their defenses under realistic and unpredictable conditions. This layer is composed by a set of modules described next.

1) Threat intelligence database: The SCE orchestrator consumes a threat intelligence database to identify the attack procedures which are applicable to a target, according to aspects like the type of operative system, the expected attack goal, the requirements for execution, among others. This database can be one included in the BAS or an external one.

2) Chaos Experiment Designer: The Chaos Experiment Designer module defines the scope and parameters of chaos-driven attack simulations. It allows users to specify requirements such as the target machine (t_i), deployment agent (a_i), number of parallel experiments (n), and security objective (o_i) (e.g., testing ransomware resilience). By integrating with threat intelligence databases, it dynamically selects adversary Tactics, Techniques, and Procedures (TTPs) that align with the

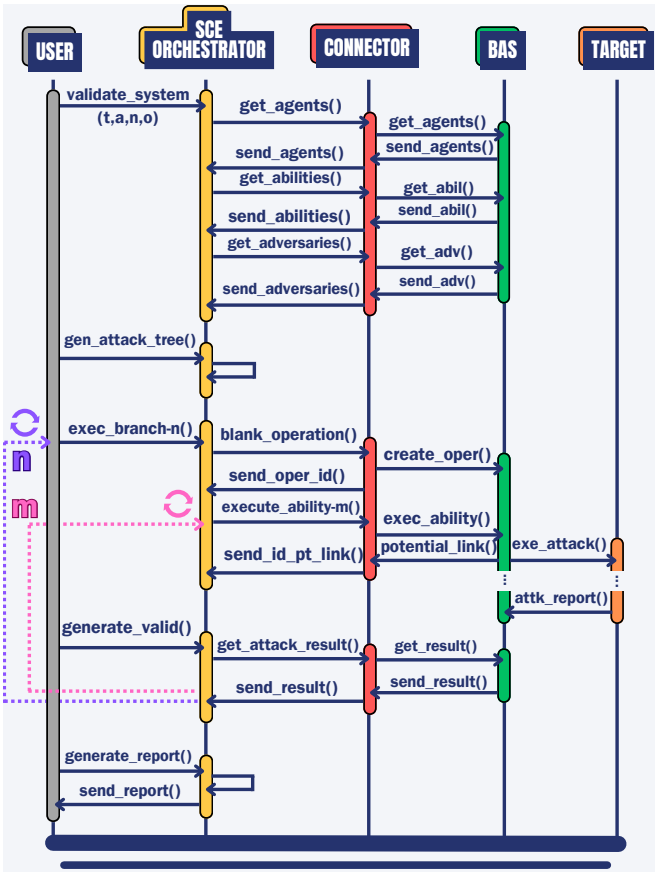


Figure 2: Flow diagram depicting the interaction between the modules of the proposed architecture

defined goals, ensuring simulations reflect real-world attack patterns.

3) Attack Tree Generator: The attack tree begins with a central attack objective, such as unauthorized access to sensitive data. From this root node, multiple branches extend, each representing a unique pathway an adversary might take to achieve the objective, based on adversary profiles sourced from MITRE Caldera. For example, one branch may involve exploiting software vulnerabilities to gain initial access, while another branch might utilize phishing techniques to obtain user credentials. Each branch further subdivides into specific tactics and techniques, aligning with the principles of Security Chaos Engineering to ensure realistic and unpredictable attack scenarios. This structured approach allows for a comprehensive simulation of diverse attack vectors, facilitating the identification and mitigation of potential security weaknesses within the target architecture.

4) ChaosXploit: We use the implementation of ChaosXploit to control the logic of SCE experiments, adapting the components needed, and taking advantage of a previously defined framework [12]. ChaosXploit is composed of modules, with two of the most important being the Exploiter and Continuous Validator. On the one hand, the Exploiter module executes attacks based on the attack tree structure, ensuring targeted actions within the simulation environment. It utilizes selected adversary tactics to perform precise attack steps, effectively challenging the security posture. On the other hand, the Continuous Validator module continuously monitors system performance and stability throughout the simulations. It tracks real-time operational parameters to detect anomalies and ensures immediate responses are initiated when deviations occur, maintaining the integrity and reliability of the experiments.

The flow between the modules of our proposed architecture is depicted in Figure 2. The *SCE Orchestrator layer* receives the user inputs (target machine (t_i), deployment agent (a_i), number of parallel experiments (n), and security objective (o_i)) and starts the validation. To achieve the validation, the *SCE Orchestrator layer* communicates with the *connector layer*, which also communicates with the *BAS layer* to obtain available agents, abilities, and adversaries. Finally, with all the obtained information (agents, abilities, adversaries), the *SCE Orchestrator layer* checks all the required components to start an attack simulation.

When the user requests the generation of an attack tree `gen_attack_tree()`, the *SCE Orchestrator layer* initiates the Attack Tree Generator module. It analyzes adversary profiles from the Threat Intelligence Database to infer potential attack paths. By leveraging predefined tactics and techniques, the generator constructs a structured attack tree, detailing each possible step based on the input parameters. The completed attack tree is then ready for execution, offering a comprehensive visual of potential threat vectors.

Upon the user's request to execute a branch `execute_branch-n()`, the *SCE Orchestrator layer* coordinates with the *BAS layer* to establish an operation. This involves executing "m" abilities designed within the attack tree branch. Each ability is processed sequentially to reflect real-world attack scenarios. The execution may involve "n" branches, depending on the complexity and objectives defined by the user. Throughout this process, the *BAS layer*'s Core Framework manages the Attack Execution Engine to ensure precise application of the designed sequences. After execution, the results are gathered and prepared for analysis in the form of an attack report `attack_report()`.

After completing the execution, the *SCE Orchestrator layer* retrieves the attack results using `get_attack_result()`. It systematically collects the outcomes and performs an initial analysis of the simulation's effectiveness. The processed results are then sent back to the user `send_result()` to validate the attack's success and effectiveness, providing insights into potential security improvements.

When the user requests to generate a report `generate_report()`, the *SCE Orchestrator layer* compiles all data and findings into a comprehensive document. This report encapsulates the simulation process, including objectives, methodologies, executed abilities, and results. Moreover, such a report highlights key vulnerabilities

discovered and suggests potential remediation steps. This detailed report serves as a critical tool for organizations to understand their current security posture and prepare strategies for future improvements.

IV. Experiments

Several experiments were created, executed, and reviewed for the validation of our proposal. All experiments are available in the project repository[1]. To ensure repeatability, Section IV-A describes the scenario, where the hardware and software elements will be detailed, Section IV-B describes the parameters needed for the simulation, Section IV-C indicates the attack tree that will be followed by the simulated adversary, Section IV-D defines the SCE experiment, and Section IV-E resumes the results of execution.

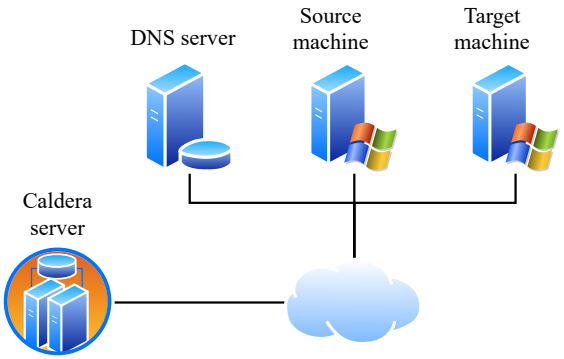


Figure 3: Network topology for the experiments

A. Settings

The components required for the experiments were virtualized through a host machine with an Arch-based distribution, AMD Ryzen 7 5800X CPU, Nvidia RTX 4070 GPU, 32GB RAM, and 2TB SSD. Four virtual machines were used in the experiments as detailed next, and following the network topology of Figure 3:

- Caldera server: This machine serves as the location where the BAS will be located and running. It is the central point where the agents report to, and from where they receive the instructions. Our proposal does not have to be run on the same machine, but on a machine which can access its API through the network. Caldera version 4.2.0. was virtualized on an Ubuntu Desktop 22.04 distribution with stock settings.
- Source machine: The source machine serves as the initial point for the cyberattack, being a machine infected by a worm that is trying to perform lateral movement to other targets within the network. It is virtualized on a Windows 10 base installation.
- Target machine: The target machine will represent our victim, which is a machine that shares a network with the source machine. It has some settings that make it vulnerable, such as a vulnerable SMB server configuration, and weak security policies.
- DNS server: A DNS server which will allow both the target and source machine to have a fully qualified domain name (FQDN), albeit a local one. This is a requirement to run certain adversary profiles and allows the communication to be done with natural names instead of IP addresses.

[1]https://github.com/noname13a/SCE-BAS

Variable name	Variable value(s)	Selected value for the experiment
t_i	$TargetList = \{Windows10_A, WindowsServer_A, UbuntuServer_A\}$	$t_1 = Windows10_A$
a_i	$AgentList = \{sandcat_A, sandcat_B, sandcat_C\}$	$a_1 = sandcat_A$
n	$n \in N^+$	1
o_i	$ObjectiveList = \{LateralMovement, Exfiltration, PrivilegeEscalataion\}$	$o_1 = LateralMovement$

Table I: Selected variables for our experiment

B. Defining the parameters of the simulation

As previously mentioned, some variables have to be determined in preparation for the simulation. First, it would be the target machine (t_i) chosen between a list of available targets (($TargetList = \{t_1, ..., t_n\}$)). Without a target, other machines found on the network could be attacked, which may or may not be the expected behavior. Setting this target ensures the attack is directed at the machine we want to test. Secondly, we must specify a deployed agent (a_i) existing in an agent list ($AgentList = \{a_1, ..., a_n\}$), which will be used to execute the abilities determined on the attack graph. Though Caldera offers us the possibility of using several agents on an operation, each with a different role, for our experiments we will limit ourselves to one agent, trying to pivot to the target machine. The last two parameters are the number of parallel experiments (n), which is a default value, and the security objective (o_i), will be chosen between a list of possible available objectives ($ObjectiveList = \{o_1, ..., o_n\}$). The selected variable values for our experiments are described in Table I.

C. Identifying the Adversary Options

Our framework composes an attack tree (Figure 4) autonomously, extracting attack goals from Caldera's adversary profiles, presented as a comprehensive final report, which serves as a foundational tool for conducting Security Chaos Engineering (SCE) experiments.

Each branch of the attack tree corresponds to one of these adversary profiles with their respective abilities depicted as nodes within each branch. This structure allows the SCE Orchestrator to utilize the attack tree in conjunction with the SCE philosophy to design chaos-driven experiments that emulate realistic adversary behaviors. For this example, we have filtered the attack tree to include adversary profiles with common attack goals of discovery, lateral movement, and execution, although some profiles possess additional abilities. This attack tree integrates multiple adversary profiles, each sharing common attack objectives such as discovery, lateral

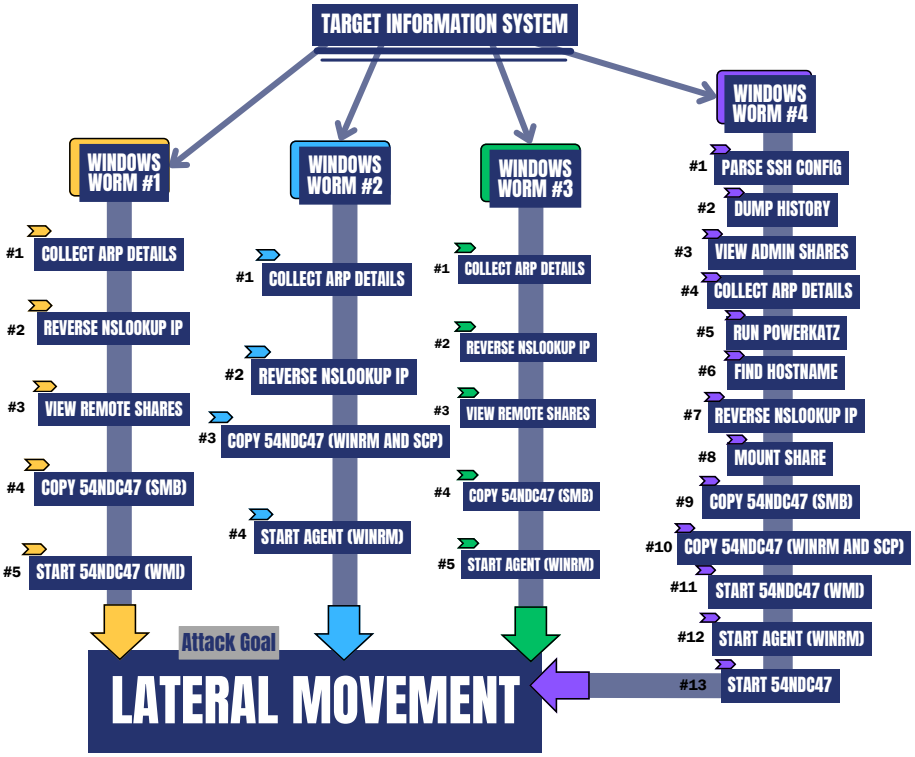


Figure 4: Attack Tree composed from adversary profiles existing in Caldera

movement, and execution. Specifically, in our experiment, we have included four distinct adversary profiles: Windows Worm #1 (SMB + WMI), Windows Worm #2 (WinRM + SCP), Windows Worm #3 (SMB + WinRM), and Worm (SMB + WinRM + WMI).

The worms described utilize various techniques for lateral movement, with some overlapping techniques. Those worms typically begin by collecting ARP details and reversing nslookup IPs to map the network environment. The movement is achieved by using protocols like SMB, WinRM, and SCP to copy and start processes on remote machines. For instance, Windows Worm #1 utilizes SMB and WMI to move laterally, while Windows Worm #2 employs WinRM and SCP. Windows Worm #3 combines both SMB for copying and WinRM for execution. The general approach involves viewing remote or admin shares, copying essential files, and starting agent processes to ensure the worm propagates effectively. The Worm adversary example illustrates a combination of these techniques, enhancing its lateral movement capabilities by parsing SSH configurations and utilizing both WMI and WinRM for process initiation, thus ensuring versatility and robustness in achieving its objectives across a network.

D. Composing SCE experiments

Our goal with this experiment stems from the question of how secure would an organization be in front of a self-replicating worm giving attackers Command and Control (C2) and Remote Access Tool (RAT) control of the computers in the network. This experiment could be performed by an organization interested in simulate cyber attacks coming from an adversary with a known arsenal, with the purpose of identify which attack vector, i.e. a branch of the tree, could be successful. Based on the attack tree presented beforehand, we can define the experiment following the SCE scientific method the following way:

- **Observability**: We have several options for observability, all related to the creation of a new agent. First, the response from Caldera for the ability execution, which we can parse to determine if it has been successful. Another would be obtaining this information through API calls, in particular, `getAgents()`, which could be used to compare it to the base call result, seeing if we have a new agent on the list.
- **Steady State**: The system in its base state, without running experiments, representing an objective that has not been attacked yet.
- **Hypothesis**: Assuming proper configuration of the services and security of the system, a lateral movement attack, like a self-replicating worm, should fail.

E. Execution of simulations

Following the execution of our experiments, we will explain first the branches that succeeded, being Windows Worm #1 and #3. Those reached the end, refuting the hypothesis mentioned in Section IV-D, and proving that the worm was

allowed to pivot to a different machine within the network.

Both worm #1 and #3 managed to implant the agent on the source machine through a shared SMB folder with incorrect permissions assigned, allowing unauthenticated users to copy files to the public directory of the machine. This file was then executed, via different methods, frist with WMIC and then with WinRM, both succeeding due to weak credentials (Worm #1) and improper configuration (Worm #3).

As for the branches that failed, starting with Windows Worm #2, this branch aborted due to not having the knowledge required to propagate the worm through SCP, this is the reason behind the failure of branch #4 as well. This failure could be mitigated with intermediate steps, which could try to bypass or find the credentials of the target machine, allowing its implantation. Once implanted, the execution methods are the same as the branches that succeeded, meaning that the only reason those branches failed was the lack of credentials. Details about the execution of experiments and the obtained results can be consulted in the project repository.

V. Conclusions and Future Work

In this paper, we introduce a framework that integrates Breach Attack Simulation (BAS) tools with Security Chaos Engineering (SCE) methodologies to conduct comprehensive security experiments across diverse target architectures. Leveraging hypotheses generated from our knowledge database and detailed attack representations, our proposal executes SCE-driven experiments to identify potential security vulnerabilities within target systems. To demonstrate the effectiveness of our proposal, we conducted a series of experiments a cyber defense scenario, assessing the security posture through SCE-based simulations. The results highlighted that some adversary profiles, i.e. some branches of the attack tree, successfully exploited the target. Our proposal is publicly available to the cybersecurity community through our the project repository[2].

Even with the results of the experiments, the proposal is still far from complete, as it requires an extensive amount of tinkering and work, and the range of scenarios it covers is fairly limited. We intend to extend the functionality to work with all of the adversary profiles provided by Caldera, but more importantly, adding the possibility of jumping from one branch to another inside the attack graph, or even taking leaps and backtracking to different nodes in it, expanding the capabilities of the solution developed.

Acknowledgment

This work has been co-funded by the European Union (project ECYSAP EYE). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Defence Fund. Neither the European Union nor the granting authority can be held responsible for them.

This work has also been partially supported by MCIN/AEI/10.13039/501100011033 NextGeneration EU/PRTR, UE, under Grant TED2021-129300B-I00, by MCIN/AEI/10.13039/501100011033/FEDER, UE, under Grant PID2021-122466OB-I00, by the Spanish National Institute of Cybersecurity (INCIBE) by the Recovery, Transformation and Resilience Plan, Next Generation EU under the strategic project DEFENDER, by the CyberDataLab (Cybersecurity and Data Science Laboratory) at the University of Murcia (Spain), and the School of Engineering, Science and Technology at the University of Rosario (Colombia).

[2]https://github.com/noname13a/SCE-BAS

References

[1] J. Kizza, *Guide to Computer Network Security*, ser. Computer Communications and Networks. Springer International Publishing, 01 2017.

[2] F. Basholli, D. Hyka, A. Basholli, A. Daberdini, and B. Mema, "Analysis of cyber-attacks through simulation," *Advanced Engineering Days (AED)*, vol. 7, pp. 120–122, 07 2023.

[3] M. Kouremetis, D. Lawrence, R. Alford, Z. Cheuvront, D. Davila, B. Geyer, T. Haigh, E. Michalak, R. Murphy, and G. Russo, "Mirage: cyber deception against autonomous cyber attacks in emulation and simulation," *Annals of Telecommunications*, vol. 79, p. 15, 03 2024.

[4] Y. Kalra, S. Upadhyay, and P. Patheja, "Advancements in Cyber Attacks and Security," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, p. 3, 02 2020.

[5] N. Mohamed, "Study of bypassing Microsoft Windows Security using the MITRE CALDERA Framework," *F1000Research*, vol. 11, p. 21, 04 2022.

[6] M. Landauer, K. Mayer, F. Skopik, M. Wurzenberger, and M. Kern, "Red Team Redemption: A Structured Comparison of Open-Source Tools for Adversary Emulation," *arXivLabs - arXiv preprint arXiv:2408.15645*, p. 12, 08 2024.

[7] K. Shortridge and A. Rinehart, *Security Chaos Engineering*. O'Reilly Media, 2023.

[8] R. M. Portase, A. Colesa, and G. Sebestyen, "SpecRep: Adversary Emulation Based on Attack Objective Specification in Heterogeneous Infrastructures," *Sensors*, p. 21, 08 2024.

[9] K. Koo, D. Moon, J.-H. Huh, S.-H. Jung, and H. Lee, "Attack Graph Generation with Machine Learning for Network Security," *Electronics*, p. 25, 04 2022.

[10] A. B. Ajmal, S. Khan, M. Alam, A. Mehbodniya, J. Webber, and A. Waheed, "Toward Effective Evaluation of Cyber Defense: Threat Based Adversary Emulation Approach," *IEEE Access*, p. 16, 05 2023.

[11] J. P. Sang Ho Oh, Jeongyoon Kim, "Dynamic Cyberattack Simulation: Integrating Improved Deep Reinforcement Learning with the MITRE-ATT&CK Framework," *Electronics*, p. 17, 07 2024.

[12] S. Palacios Chavarro, P. Nespoli, D. Díaz-López, and Y. Niño Roa, "On the Way to Automatic Exploitation of Vulnerabilities and Validation of Systems Security through Security Chaos Engineering," *Big Data and Cognitive Computing*, vol. 7, no. 1, 2023.

[13] M. Bedoya, S. Palacios, D. Díaz-López, E. Laverde, and P. Nespoli, "Enhancing DevSecOps practice with Large Language Models and Security Chaos Engineering," *International Journal of Information Security*, vol. 23, no. 6, pp. 3765–3788, Dec 2024.

Sesión Formación I

Laboratorio de aprendizaje de Inyecciones SQL en entornos de Hacking Ético

1st Alejandro Bofill González
Grupo de Ingeniería de Medios
Universidad de Extremadura
abofillg@alumnos.unex.es

2nd David Gordillo Burrero
Grupo de Ingeniería de Medios
Universidad de Extremadura
dgordillj@alumnos.unex.es

3rd Alberto López Trigo
Grupo de Ingeniería de Medios
Universidad de Extremadura
albertotl@unex.es

4th Andrés Caro Lindo
Grupo de Ingeniería de Medios
Universidad de Extremadura
andresc@unex.es

***Resumen*—Las inyecciones SQL (SQLi) representan una de las vulnerabilidades más críticas en aplicaciones web, permitiendo a atacantes manipular bases de datos mediante la inserción de código malicioso. En este trabajo, se presenta un laboratorio didáctico diseñado para facilitar el aprendizaje sobre el funcionamiento de las inyecciones SQL a nivel básico y avanzado. El laboratorio está orientado a usuarios con conocimientos limitados en ciberseguridad, brindando una experiencia accesible para comprender desde los conceptos fundamentales hasta la ejecución y mitigación de ataques SQLi.**

Se proporciona un entorno controlado en el que se simulan ataques SQLi en bases de datos *Oracle* y *PostgreSQL*, abarcando diversas técnicas como inyecciones basadas en errores, *Union Attack*, *Boolean-Based SQL Injection* y *Blind SQL Injection*. A través de ejercicios prácticos, los participantes pueden experimentar de primera mano cómo estas vulnerabilidades pueden ser explotadas, qué condiciones las hacen posibles y cuáles son las mejores estrategias de defensa, como el uso de consultas parametrizadas y la validación estricta de entradas.

Los resultados obtenidos reflejan la importancia de un enfoque práctico para el aprendizaje de la seguridad en bases de datos, proporcionando un recurso accesible y educativo que permite a los usuarios desarrollar habilidades esenciales para la identificación y mitigación de ataques SQLi en entornos reales.

***Index Terms*—SQL Injection, Hacking Ético, Ciberseguridad**

I. Introducción

Las inyecciones SQL (SQLi) han sido identificadas como una de las principales vulnerabilidades en aplicaciones web, permitiendo a actores malintencionados manipular bases de datos mediante la inserción de código SQL no autorizado. Este tipo de ataques ha sido ampliamente documentado y sigue figurando entre las principales amenazas de seguridad según el *Open Web Application Security Project* (OWASP [5]).

La persistencia de las inyecciones SQL como amenaza se evidencia en incidentes recientes. En septiembre de 2024, investigadores de seguridad descubrieron una vulnerabilidad en un sitio web de terceros utilizado por aerolíneas para acceder al sistema *Known Crewmember* (KCM) de la Administración de Seguridad en el Transporte (TSA) de Estados Unidos [2]. Esta vulnerabilidad permitía, mediante técnicas básicas de inyección SQL, añadir individuos no autorizados como miembros de la tripulación, potencialmente eludiendo los controles de seguridad y accediendo a áreas restringidas.

La relevancia de las inyecciones SQL radica en su capacidad para comprometer la integridad, confidencialidad y disponibilidad de los datos almacenados en sistemas de gestión de bases de datos. La explotación de estas vulnerabilidades puede resultar en el acceso no autorizado a información sensible, la modificación o eliminación de datos, e incluso la toma de control total sobre los servidores de base de datos afectados.

A pesar de la existencia de múltiples estrategias de mitigación, como el uso de consultas parametrizadas y la validación estricta de entradas, muchas aplicaciones aún presentan fallos de seguridad que pueden ser explotados mediante SQLi. La razón principal detrás de esta persistencia radica en la falta de formación y conciencia sobre los principios fundamentales de seguridad en el desarrollo de software.

Con el objetivo de abordar esta problemática, este trabajo presenta un laboratorio didáctico diseñado para facilitar el aprendizaje práctico sobre inyecciones SQL. A través de un entorno controlado, se permite a los participantes experimentar con diferentes tipos de ataques SQLi y comprender sus mecanismos a bajo nivel. De esta manera, se busca no solo demostrar la peligrosidad de estas vulnerabilidades, sino también proporcionar las herramientas necesarias para su detección y prevención en aplicaciones reales.

El laboratorio propuesto abarca una serie de ejercicios estructurados en torno a diferentes técnicas de inyección SQL, incluyendo inyecciones basadas en errores, *Union Attack*, inyecciones booleanas y *Blind SQL Injection*. Cada ejercicio ha sido diseñado para ilustrar un aspecto particular de la vulnerabilidad y fomentar el desarrollo de buenas prácticas en el desarrollo de software seguro.

Este proyecto está dirigido a estudiantes y profesionales que cuenten, al menos, con conocimientos básicos de programación y nociones elementales de SQL, sin que sea necesaria experiencia previa en ciberseguridad. Su diseño accesible permite una amplia variedad de usos formativos: desde demostraciones en clase por parte del docente, hasta

prácticas guiadas en laboratorio o sesiones de autoestudio. El entorno se ejecuta de forma local y autónoma, lo que facilita su integración en contextos educativos con recursos limitados, garantizando una experiencia práctica, segura, controlada y reproducible para el aprendizaje progresivo de las inyecciones SQL.

El repositorio del proyecto está disponible públicamente en GitHub bajo la licencia *"Creative Commons Attribution-NonCommercial 4.0 International"* [14]. Puede descargarse localmente y utilizarse de forma sencilla.

II. Estado del Arte

En el ámbito de la ciberseguridad, la formación práctica en técnicas de inyección SQL (SQLi) es esencial para comprender y mitigar eficazmente estas vulnerabilidades. Diversas plataformas han sido desarrolladas con este propósito, cada una ofreciendo enfoques y características particulares. A continuación, se presenta un análisis técnico y formal de las principales soluciones disponibles, destacando sus ventajas y limitaciones:

- **PortSwigger Web Security Academy** [10]: Ofrece laboratorios interactivos que abarcan una amplia gama de vulnerabilidades, incluyendo inyecciones SQL, con guías detalladas y entornos prácticos.
- **Udemy** [16]: Plataforma de aprendizaje en línea con una variedad de cursos sobre inyección SQL y hacking ético, impartidos por diferentes instructores.
- **Backtrack Academy** [2]: Especializada en cursos de seguridad informática, ofreciendo módulos dedicados a la inyección SQL con un enfoque práctico y detallado.
- **Hack The Box** [1]: Plataforma que ofrece laboratorios virtuales y desafíos prácticos orientados al hacking ético y a la explotación de vulnerabilidades.

Se presenta una tabla comparativa [I] que resume las ventajas y limitaciones de cada plataforma:

Ventaja Distintiva del Laboratorio Propuesto

El laboratorio propuesto se distingue por ofrecer una característica única: la visualización en tiempo real del código SQL que se ejecuta en segundo plano durante las prácticas de inyección. Esta funcionalidad aporta múltiples beneficios:

- Transparencia total: Los participantes pueden observar directamente cómo sus interacciones afectan la base de datos, facilitando una comprensión más profunda de las operaciones internas.
- Aprendizaje reforzado: Al correlacionar las acciones realizadas con el código ejecutado, se promueve una asimilación más efectiva de los conceptos y técnicas de inyección SQL.
- Detección de errores: La posibilidad de visualizar el código en ejecución permite identificar y corregir rápidamente posibles fallos o malentendidos en la construcción de las consultas.

Esta capacidad de monitoreo en tiempo real del código ejecutado no es común en las plataformas mencionadas anteriormente, posicionando al laboratorio propuesto como una herramienta innovadora y eficaz para el aprendizaje profundo de las inyecciones SQL.

Cuadro I
Comparativa de Plataformas de Aprendizaje en Inyección SQL

Plataforma	Ventajas	Limitaciones
PortSwigger Web Security Academy	▪ Laboratorios actualizados que reflejan vulnerabilidades contemporáneas. ▪ Acceso gratuito a recursos educativos de alta calidad.	▪ Laboratorios de caja negra en los que no se puede analizar el código que se ejecuta por detrás a la hora de hacer una inyección.
Udemy	▪ Amplia variedad de cursos, adaptados a distintos niveles de conocimiento. ▪ Acceso de por vida al contenido del curso adquirido.	▪ Calidad variable según el instructor. ▪ Muchos cursos son de pago, lo que puede ser una limitación.
Backtrack Academy	▪ Contenido diseñado para la comunidad hispanohablante. ▪ Enfoque práctico con laboratorios que simulan escenarios reales.	▪ Acceso limitado sin suscripción o pago único.
Hack The Box	▪ Entorno virtual con una amplia variedad de desafíos y máquinas vulnerables. ▪ Comunidad activa que fomenta el aprendizaje colaborativo. ▪ Cubre múltiples vectores de ataque.	▪ Requiere suscripción de pago para acceder a contenidos avanzados. ▪ Laboratorios centrados en la explotación y no en entender cómo ocurre una inyección.

III. Método

El presente estudio se fundamenta en el desarrollo de un laboratorio interactivo diseñado para el análisis y comprensión de diversas técnicas de inyección SQL. Con el propósito de proporcionar un entorno de aprendizaje práctico y estructurado, se han implementado cinco laboratorios, cada uno orientado a la exploración de diferentes modalidades de ataques SQLi. La arquitectura del laboratorio se ha construido utilizando bases de datos *Oracle Database* y *PostgreSQL Database*, así como *Python* para la gestión de la ejecución de consultas y la construcción de una interfaz web que muestra el código ejecutado en tiempo real.

El laboratorio se encuentra diseñado para inicializarse de manera autónoma mediante la ejecución de una serie de *scripts* que crean las estructuras de datos necesarias en ambas bases de datos, asegurando así la coherencia y reproducibilidad de los experimentos. Este procedimiento automatizado garantiza que cada instancia del laboratorio disponga de las tablas y registros adecuados para la ejecución de los diferentes ejercicios prácticos.

Cada laboratorio se compone de los siguientes elementos:

- Una descripción detallada de la técnica de inyección SQL correspondiente, proporcionando el contexto teórico necesario para su comprensión.
- Un panel de autenticación que simula un sistema vulnerable, permitiendo la experimentación con técnicas de inyección SQL en un entorno controlado.
- Un desplegable que facilita la selección y copia de cadenas de inyección SQL predefinidas, agilizando la ejecución de pruebas sin necesidad de formular manualmente las consultas.
- Un visor en tiempo real que muestra la consulta SQL generada en el sistema tras cada intento de autenticación, permitiendo analizar el impacto de la inyección de código sobre la consulta ejecutada.

El diseño modular del laboratorio permite la adaptación progresiva de los ataques, modificando la lógica subyacente del código ejecutado para abordar diferentes enfoques de explotación. Las técnicas de inyección SQL implementadas en los cinco laboratorios incluyen:

- Inyección basada en errores, permitiendo la obtención de información sensible a partir de los mensajes de error de la base de datos.
- Inyección de tipo *Union Attack*, que posibilita la recuperación de datos de otras tablas mediante la combinación de resultados.
- Inyecciones basadas en booleanos, explotando consultas condicionales para inferir información sin generar errores visibles.
- Inyecciones a ciegas basadas en booleanos, evaluando respuestas binarias para reconstruir la estructura de la base de datos.
- Inyecciones a ciegas basadas en tiempo, utilizando retardos en la ejecución de consultas para extraer información de forma encubierta.

Cada uno de estos enfoques ha sido implementado con ligeras variaciones en el código ejecutado, permitiendo analizar el impacto de diferentes estrategias de protección y validación de datos. En conjunto, este laboratorio constituye una herramienta integral para la formación en seguridad informática, proporcionando un entorno controlado donde los usuarios pueden experimentar y comprender las implicaciones prácticas de las inyecciones SQL y estrategias de mitigación.

III-A. Laboratorio 1: Inyección Basada en Errores

La inyección SQL basada en errores se basa en la explotación de mensajes detallados que revelan datos sensibles sobre el esquema, los nombres de tablas y columnas, e incluso valores almacenados en la base de datos.

III-A1. Mecanismo del Ataque: La explotación de errores en bases de datos se fundamenta en la falta de control sobre los mensajes generados por el SGBD cuando se ejecutan consultas mal formadas. El atacante puede aprovechar esta información para refinar sus consultas y obtener acceso no autorizado.

III-A2. Ejemplo de Inyección Basada en Errores: Se presenta un ejemplo de una inyección SQL basada en errores, donde el atacante intenta obtener información sobre la existencia de tablas en la base de datos:

```
SELECT * FROM Usuarios WHERE username = ''
OR 1=(SELECT * FROM tabla_inexistente)
-- ' AND password = 'cualquier_input'
```

Esta consulta provoca un error si la tabla mencionada no existe, lo que puede revelar detalles sobre la estructura de la base de datos y su esquema interno.

III-A3. Inyecciones Disponibles en el Laboratorio: Para ilustrar esta técnica, el laboratorio permite probar diferentes consultas predefinidas que explotan la generación de errores para extraer información específica:

- Obtención de información sobre las tablas disponibles en la base de datos.
- Extracción de metadatos sobre las columnas de una tabla específica.
- Identificación del tipo de SGBD en uso a partir de mensajes de error.

III-B. Laboratorio 2: Inyección Basada en Union Attack

La inyección SQL basada en *UNION* es una técnica en la que un atacante aprovecha la cláusula `UNION SELECT` para combinar los resultados de una consulta legítima con datos maliciosamente solicitados. Este método permite extraer información confidencial de la base de datos, incluyendo nombres de tablas, columnas y registros sensibles. Para ejecutar este ataque, el atacante debe identificar vulnerabilidades en la aplicación web, determinar la cantidad de columnas presentes en la consulta original y posteriormente construir una consulta manipulada que utilice `UNION SELECT` para obtener los datos deseados.

III-B1. Mecanismo del Ataque: El ataque mediante *Union Injection* se lleva a cabo en tres fases principales:

- **Identificación de puntos vulnerables**: Se buscan campos de entrada en la aplicación web donde las consultas SQL no están protegidas contra inyecciones. Usualmente, los formularios de inicio de sesión, parámetros en URL y cuadros de búsqueda son susceptibles a este tipo de ataque.
- **Determinación del número de columnas**: Para construir una consulta válida con `UNION SELECT`, el atacante necesita conocer la cantidad de columnas que la consulta original devuelve. Para ello, se utilizan técnicas como `ORDER BY` para incrementar el número de columnas en la consulta hasta que se genere un error o consultas de prueba con `UNION SELECT NULL` para ajustar la cantidad de columnas.
- **Construcción de la inyección**: Una vez determinado el número de columnas, el atacante elabora una consulta manipulada con `UNION SELECT` que le permita acceder a información confidencial almacenada en la base de datos. Dependiendo del SGBD, se pueden extraer nombres de bases de datos, versiones, nombres de tablas y registros sensibles.

III-B2. Ejemplo de Inyección Basada en Union Attack: A continuación, se muestra un ejemplo de inyección SQL utilizando la cláusula `UNION SELECT` para obtener el nombre de la base de datos en un entorno *Oracle*:

```
cualquier_input' UNION SELECT 1,
    ora_database_name, NULL AS
    nombre_bd_relleno_1,
NULL AS nombre_bd_relleno_2 FROM dual --
```

En este caso, el atacante fuerza la ejecución de una consulta adicional que recupera el nombre de la base de datos, rellenando las columnas restantes con valores `NULL` para garantizar la compatibilidad con la consulta original.

III-B3. Inyecciones Disponibles en el Laboratorio: Se propone un conjunto de inyecciones SQL predefinidas para analizar diferentes aspectos del ataque basado en *Union Injection*:

- Obtención del nombre de la base de datos.
- Obtención de la versión del sistema de gestión de bases de datos.
- Enumeración de todas las tablas disponibles en la base de datos.
- Enumeración de todas las columnas dentro de una tabla específica.
- Obtención de registros de una tabla de interés.

III-C. Laboratorio 3: Inyección Basada en Booleanos

La inyección SQL basada en booleanos es una técnica mediante la cual un atacante explota vulnerabilidades en las consultas SQL de una aplicación para inferir información sobre la base de datos. Este método se fundamenta en el análisis de las respuestas de la aplicación a consultas booleanas, es decir, aquellas que evalúan condiciones como verdaderas o falsas. A diferencia de otras técnicas, este ataque no requiere acceso directo a los datos almacenados en la base de datos; en su lugar, se utiliza la variación en las respuestas de la aplicación (por ejemplo, cambios en el contenido, mensajes de error o tiempos de carga) para deducir información sensible. La mitigación de este tipo de ataque requiere una rigurosa validación de las entradas, el empleo de consultas parametrizadas y una adecuada configuración de los mensajes de error.

III-C1. Mecanismo del Ataque: El proceso de explotación de la inyección basada en booleanos se estructura en las siguientes fases:

- **Identificación de puntos vulnerables**: Se examinan los campos de entrada de la aplicación web para detectar aquellos que no cuentan con una sanitización adecuada, lo que permite la inyección de código SQL.
- **Construcción de consultas booleanas**: Una vez identificados los puntos vulnerables, se elaboran consultas que evalúan condiciones lógicas. El atacante observa las diferencias en las respuestas de la aplicación, determinando si la condición evaluada es verdadera o falsa.
- **Inferencia de datos sensibles**: Tras confirmar la vulnerabilidad, el atacante utiliza consultas booleanas específicas para deducir información sobre la estructura y el contenido de la base de datos, basándose en las variaciones en las respuestas.

III-C2. Ejemplo de Inyección: Obtención del Número de Caracteres: A continuación, se presenta un ejemplo de inyección SQL que permite determinar si el número de caracteres del campo `username` es igual a 5. Esta consulta evalúa una condición booleana y, en función del resultado, provoca un error (división por cero) o se ejecuta correctamente:

```
' OR ( SELECT CASE WHEN ( LENGTH(username) = 5
    )
THEN 1/0 ELSE 1 END FROM
( SELECT username, ROWNUM AS rn FROM Usuarios
    )
WHERE rn = 1) = 1 --
```

En este ejemplo, la inyección evalúa si la longitud del campo `username` es igual a 5. Si la condición se cumple, se provoca una división por cero (1/0), lo que genera un error detectable en la respuesta de la aplicación; de lo contrario, se retorna 1, permitiendo al atacante inferir la longitud del campo.

III-C3. Inyecciones Propuestas: El laboratorio incorpora una serie de inyecciones predefinidas para demostrar la eficacia de este tipo de ataque. Entre las inyecciones disponibles se incluyen:

- Obtención del número de caracteres de un campo.
- Obtención de un carácter específico de un campo.

III-D. Laboratorio 4: Inyección Blind basada en Condiciones Booleanas

Las ***Blind SQL Injections*** son un tipo de ataque en el que el atacante puede inferir información sobre la base de datos, aun cuando la aplicación no proporciona mensajes de error explícitos o respuestas directas. A diferencia de las inyecciones SQL tradicionales, donde los resultados de las consultas maliciosas pueden ser visibles en la interfaz de la aplicación, las inyecciones ***Blind*** se basan en observar diferencias en el comportamiento de la aplicación para deducir información sensible. Este método es especialmente efectivo en aplicaciones que han implementado medidas básicas de seguridad, como la supresión de mensajes de error, pero que siguen siendo vulnerables debido a la falta de validación adecuada de las entradas del usuario.

Dado que las inyecciones ***Blind*** no reportan mensajes de error internos, el atacante debe buscar otros medios para obtener información. Una técnica común consiste en manipular la cookie de sesión, aprovechando que la mayoría de las aplicaciones web utilizan cookies para almacenar información de sesión (por ejemplo, el ID de usuario o el *token* de autenticación). Al modificar manualmente la *cookie*, se pueden obtener indicios sobre datos sensibles. Además, este método utiliza consultas con condiciones booleanas (TRUE o FALSE) para observar cómo varía el comportamiento de la aplicación;

por ejemplo, enviando una consulta como `' OR '1'='1` para verificar si la condición se evalúa como verdadera.

III-D1. Mecanismo del Ataque: El proceso de explotación en este laboratorio se estructura en las siguientes fases:

- **Identificación de puntos vulnerables**: Se examinan los campos de entrada de la aplicación para detectar aquellos que no filtran adecuadamente las entradas del usuario, permitiendo la inyección de código SQL.
- **Construcción de consultas booleanas**: Se elaboran consultas que evalúan condiciones lógicas. El atacante observa las diferencias en el comportamiento de la aplicación (por ejemplo, variaciones en el tiempo de carga o en la respuesta) para determinar si la condición se cumple o no.
- **Inferencia de datos sensibles**: A partir de las respuestas obtenidas, el atacante deduce información sobre la estructura y el contenido de la base de datos, como la longitud de un campo o el contenido de un carácter específico.

III-D2. Ejemplo de Inyección: Obtención de la Longitud de un Campo: A continuación, se presentan dos payloads utilizados para inferir la longitud del campo `username`. Se deben ejecutar dos consultas: una para el caso en que la condición se evalúa como verdadera y otra para el caso en que se evalúa como falsa.

a) Payload para el caso TRUE

```
d382yd8n21df4314fn817yf6834188ls023d8d ' AND (
     SELECT CASE WHEN ( LENGTH(username) = 5 )
THEN 1 ELSE 1/0 END FROM ( SELECT username,
    ROWNUM AS rn FROM Usuarios )
WHERE rn = 1) = 1 --
```

b) Payload para el caso FALSE

```
d382yd8n21df4314fn817yf6834188ls023d8d ' AND (
     SELECT CASE WHEN ( LENGTH(username) = 33
    )
THEN 1 ELSE 1/0 END FROM ( SELECT username,
    ROWNUM AS rn FROM Usuarios )
WHERE rn = 1) = 1 --
```

III-D3. Inyecciones Propuestas: El laboratorio incluye inyecciones predefinidas para profundizar en este tipo de ataque:

- Obtención de la longitud de un campo de la base de datos.
- Obtención de un carácter específico de un campo de la base de datos.

III-E. Laboratorio 5: Inyección Blind basada en Tiempo

La inyección SQL basada en tiempo es una técnica utilizada para explotar vulnerabilidades en aplicaciones que interactúan con bases de datos. Este tipo de ataque se inscribe dentro de la categoría de inyección SQL a ciegas y se caracteriza por depender del tiempo de respuesta del servidor para deducir información sensible. El atacante manipula consultas SQL para introducir deliberados retrasos en la respuesta, lo que permite inferir la veracidad de ciertas condiciones basadas en el tiempo de espera observado.

Destacar que este laboratorio solamente está implementado en PostgreSQL debido a la imposibilidad existente en Oracle de ejecutar funciones que paralicen el hilo del programa en una sentencia SQL.

III-E1. Mecanismo del Ataque: El mecanismo de la inyección blind basada en tiempo se estructura en las siguientes fases:

- **Manipulación de consultas SQL:** El atacante envía entradas maliciosas que incluyen comandos SQL diseñados para retrasar intencionadamente la respuesta del servidor.
- **Uso de funciones de espera:** Se emplean funciones específicas del SGBD para generar retrasos controlados. Por ejemplo, en PostgreSQL se utiliza la función `pg_sleep(n)`, que pausa la ejecución por `n` segundos.
- **Interpretación del tiempo de respuesta:** Mediante el análisis del tiempo que tarda el servidor en responder, el atacante puede deducir si la condición evaluada es verdadera o falsa, permitiéndole inferir información sensible, como el contenido de campos o la existencia de determinados datos.

III-E2. Ejemplo de Inyección: Obtención del Primer Carácter del Nombre de la Base de Datos: A continuación se muestra un ejemplo de inyección SQL que utiliza la función `pg_sleep` para inferir el primer carácter del nombre de la base de datos. En este caso, la consulta provoca un retraso de 5 segundos si el primer carácter es igual a `'n'` y un retraso de 0 segundos en caso contrario:

```
SELECT * FROM usuarios
WHERE nombre = '' OR
( CASE WHEN SUBSTRING((SELECT current_database
    ()), 1, 1) = 'n'
THEN pg_sleep(5)
ELSE pg_sleep(0) END ) --
AND contrasena = '';
```

III-E3. Inyecciones Propuestas: El laboratorio también ofrece un conjunto de inyecciones predefinidas para explorar diferentes vectores de ataque mediante inyección basada en tiempo, entre las cuales se incluyen:

- Inyección para obtener “nombres de usuario”.
- Inyección para obtener “nombres de la base de datos”.
- Inyección para obtener “usuarios y contraseñas concatenados”.

IV. Resultados y Discusión

En esta sección se presentan los resultados obtenidos tras la ejecución de los laboratorios, analizando el impacto de las diferentes técnicas de inyección SQL y su mitigación. Se incluyen capturas de pantalla y fragmentos de código relevantes para ilustrar los hallazgos.



Figura 1. Interfaz del laboratorio de inyección SQL

La Figura 1 muestra la interfaz del laboratorio, que incluye un apartado por cada tipo de inyección SQL. Cada sección contiene una descripción detallada de la técnica, un panel de autenticación, un desplegable con consultas predefinidas y un visor en tiempo real que muestra la consulta SQL generada en el sistema tras cada intento de autenticación. Esta estructura facilita la experimentación y el análisis de los resultados obtenidos en cada laboratorio.

IV-A. Inyección Basada en Errores

Este tipo de inyección SQL aprovecha los mensajes de error generados por la base de datos para obtener información sensible. En los experimentos realizados, se logró obtener detalles sobre la estructura de las tablas y los nombres de columnas a partir de los errores devueltos por el sistema.

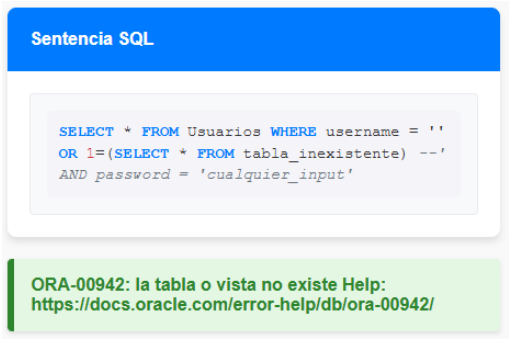


Figura 2. Exposición de información sobre nombre de tablas

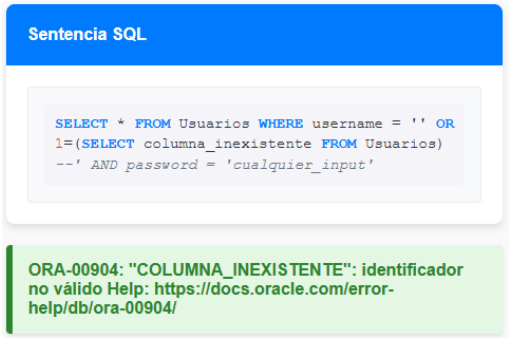


Figura 3. Exposición de información sobre nombre de columnas

En la Figura 2, se muestra cómo un mensaje de error revela información sobre el nombre de tablas en la base de datos, confirmando que la tabla *"tabla_inexistente"* no existe. De igual forma, en la Figura 3, se observa como *"columna_inexistente"* no es una columna válida, lo que permite inferir los nombres de las columnas existentes mediante un ataque de fuerza bruta, por ejemplo.

IV-B. Inyección Union Attack

Esta técnica permite recuperar datos de otras tablas al unirlas con la consulta original. Mediante una de las inyecciones implementadas se logra extraer información sobre el nombre de todas las tablas presentes en la base de datos.

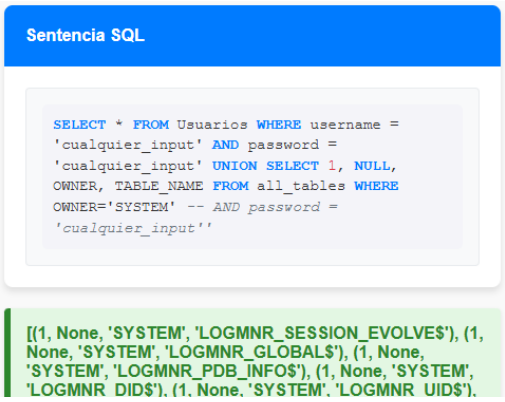


Figura 4. Uso de *Union Attack* para extraer datos de tablas adicionales.

En la Figura 4, se muestra cómo este tipo de inyección permite recuperar información de tablas adicionales, en este caso, todas las tablas de la base de datos. La consulta original se modifica para incluir una unión con *"all_tables"*, lo que resulta en la exposición del nombre de todas las tablas de la base de datos.

IV-C. Inyección Basada en Booleanos

Este método explota consultas condicionales para inferir información sobre la base de datos sin generar errores visibles. En el laboratorio, se lograron determinar valores booleanos que revelaban si una condición específica era verdadera o falsa.

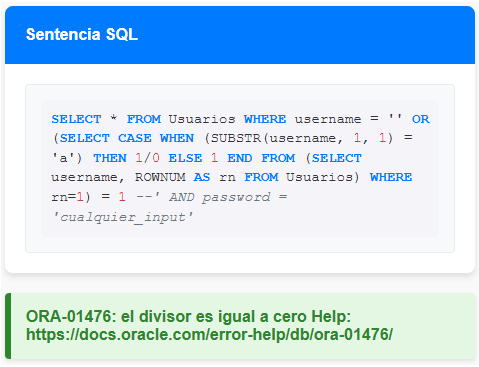


Figura 5. Ejemplo de inyección SQL basada en booleanos.

En la Figura 5, se muestra cómo la inyección basada en booleanos permite inferir información sobre la base de datos mediante consultas condicionales. En este caso, se evalúa si el primer caracter de la primera tupla de la tabla *"usuarios"* es *"a"*. En ese caso, la consulta ejecuta la casuística *1/0*, lo que provoca un error de división por cero. De esta manera se puede inferir uno a uno los caracteres de cada tupla.

IV-D. Inyección Blind Basada en Tiempo

Esta técnica introduce retardos en la ejecución de consultas para exfiltrar información. Durante las pruebas, se demostró cómo los tiempos de respuesta variaban en función de la información solicitada.

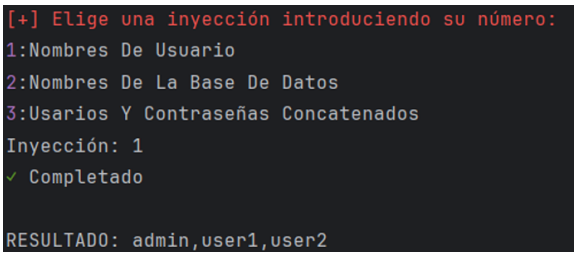


Figura 6. Uso de inyección SQL basada en tiempo.

En la Figura 6, se muestra el resultado de la ejecución de un *script* que introduce retardos en la consulta para inferir información sobre la base de datos. Si la respuesta demora más de lo indicado, se puede inferir que la condición es verdadera, permitiendo extraer información de cada tupla de la tabla. Para la explotación de este tipo de inyección se desarrolló un *script* el cual realizaba una fuerza bruta a cada caracter de un campo e infería si era correcto o no en base al tiempo de respuesta.

IV-E. Inyección Blind Basada en Condiciones Booleanas

En este caso, se realizan consultas condicionales diseñadas para extraer información sin generar ningún error visible ni retrasos apreciables. A través de múltiples peticiones, se puede reconstruir información de la base de datos.

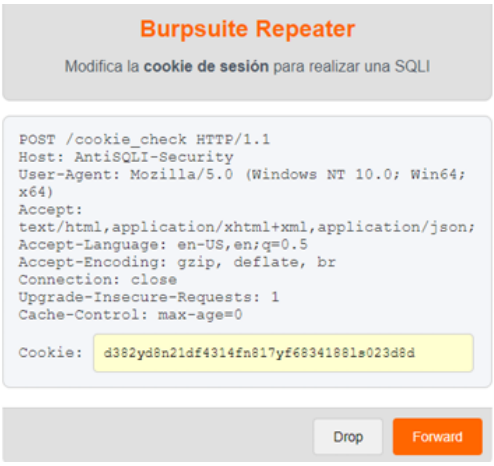


Figura 7. Inyección SQL *blind* basada en booleanos.

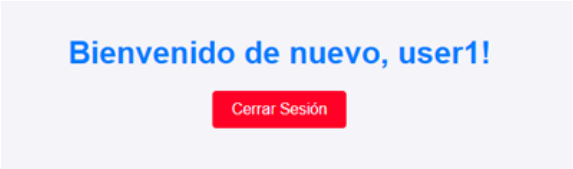


Figura 8. Confirmación de la correcta casuística.

En las Figuras 7 y 8, se muestra cómo la inyección *blind* basada en booleanos permite inferir información sobre la base de datos mediante consultas condicionales. Al tratarse de inyecciones ***blind***, se simuló la intercepción de la captura de petición *HTTP* al formulario de inicio de sesión, la cual tenía un parámetro ***cookie***. En este parametro se realizaría la inyección, de manera similar a las inyecciones basadas en booleanos. En caso de cumplirse la condición, se redirige a la página que se puede observar en la Figura 8, infiriendo de esta manera si la condición es verdadera o falsa. Al igual que en la inyección basada en tiempo, se realizó un *script* que probaba si cada caracter de un campo determinando si era correcto o no en base a la respuesta de la página., ya que sería la manera más rápida y eficiente de explotar este tipo de inyección.

IV-F. Aplicación Docente y Evaluación del Aprendizaje

Desde el punto de vista docente, la evaluación del uso del laboratorio plantea ciertos retos derivados de su ejecución local, que impide disponer de un sistema centralizado para registrar la interacción de los estudiantes. Sin embargo, su diseño ofrece diversas oportunidades para plantear actividades evaluables de forma efectiva. Una de las estrategias más recomendables consiste en pedir al alumnado que elabore inyecciones SQL propias, distintas de las consultas predefinidas que ya incorpora el entorno, con el objetivo de explorar nuevas formas de explotación y comprobar la comprensión real de los mecanismos subyacentes.

Este tipo de propuesta fomenta el pensamiento crítico y permite evaluar el nivel de asimilación de los contenidos trabajados en el laboratorio. Además, puede complementarse con evidencias como capturas de pantalla, descripciones escritas del proceso o grabaciones breves donde el estudiante explique las decisiones tomadas. A pesar de no contar con métricas automáticas, este enfoque cualitativo permite una valoración eficaz del aprendizaje, adaptable tanto a contextos presenciales como remotos.

Del mismo modo, se espera que los usuarios que utilicen el laboratorio en sesiones de autoestudio adopten una actitud activa y exploratoria, intentando construir nuevas inyecciones más allá de las propuestas iniciales.

IV-G. Discusión

Los resultados obtenidos muestran que cada técnica de inyección SQL tiene un impacto distinto en la seguridad de la base de datos. En particular:

- La inyección basada en errores es efectiva en entornos donde los mensajes de error no han sido suprimidos.

- El *Union Attack* permite la extracción masiva de datos si las consultas no están correctamente protegidas.
- Las inyecciones booleanas permiten aprovechar consultas condicionales para inferir información sensible.
- Las inyecciones *blind* basadas en booleanos y tiempo pueden ser efectivas en sistemas que no presentan errores directos, aunque son más lentas en ejecución.

V. Conclusiones

En este trabajo se ha desarrollado un laboratorio didáctico diseñado para el aprendizaje práctico sobre las inyecciones SQL (SQLi), abordando sus diferentes modalidades y técnicas de explotación. A través de la implementación de un entorno interactivo basado en bases de datos *Oracle* y *PostgreSQL*, se permite experimentar de manera controlada con ataques como inyecciones basadas en errores, *Union Attack*, inyecciones booleanas y *Blind SQL Injection*.

El laboratorio propuesto permite observar en tiempo real el impacto de las inyecciones SQL, facilitando una comprensión profunda y práctica de estas vulnerabilidades y sus mitigaciones. Este enfoque educativo fortalece la preparación en ciberseguridad, cumpliendo con el objetivo propuesto de solventar la falta de formación y conciencia sobre los principios fundamentales de seguridad en el desarrollo de software.

Agradecimientos

Esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation) - Instituto Nacional de Ciberseguridad (INCIBE) en el proyecto C109/23 ”Proyecto estratégico UEx (Escuela Politécnica de Cáceres)-INCIBE”

Referencias

[1] Backtrack Academy. Disponible en: https://backtrackacademy.com
[2] Bypassing airport security via SQL injection. Disponible en https://ian.sh/tsa
[3] Hack The Box. Disponible en: https://www.hackthebox.com
[4] Oracle, “Oracle Database SQL Language Reference”. Disponible en: https://docs.oracle.com/en/database/oracle/oracle-database/19/sqlrf/index.html
[5] OWASP Top 10. Disponible en: https://owasp.org/Top10/
[6] OWASP, “Blind SQL Injection”. Disponible en: https://owasp.org/www-community/attacks/Blind_SQL_Injection
[7] OWASP, “SQL Injection”. Disponible en: https://owasp.org/www-community/attacks/SQL_Injection
[8] OWASP, “SQL Injection Prevention Cheat Sheet”. Disponible en: https://cheatsheetseries.owasp.org/cheatsheets/SQL_Injection_Prevention_Cheat_Sheet.html
[9] PostgreSQL, “PostgreSQL Documentation”. Disponible en: https://www.postgresql.org/docs/
[10] PortSwigger Academy. Disponible en: https://portswigger.net
[11] PortSwigger, “Burp Suite Documentation”. Disponible en: https://portswigger.net/burp/documentation
[12] PortSwigger, “SQL Injection”. Disponible en: https://portswigger.net/web-security/sql-injection
[13] Python, “Python Documentation”. Disponible en: https://docs.python.org/3/
[14] Repositorio público del proyecto en GitHub: https://github.com/gordillo7/SQLiLab
[15] Stack Overflow. Disponible en: https://stackoverflow.com/
[16] Udemy. Disponible en: https://www.udemy.com
[17] W3Schools, “SQL Tutorial”. Disponible en: https://www.w3schools.com/sql/

De novatos a expertos: Creando una comunidad estudiantil para fomentar la ciberseguridad

Agustín Javier Di Bartolo
Universidad de Extremadura
Escuela Politécnica de Cáceres
adibartolo@unex.es

Alberto López Trigo
Universidad de Extremadura
Escuela Politécnica de Cáceres
albertolt@unex.es

David Gordillo Burrero
Universidad de Extremadura
Escuela Politécnica de Cáceres
dgordillj@alumnos.unex.es

Alejandro Bofill González
Universidad de Extremadura
Escuela Politécnica de Cáceres
abofillg@alumnos.unex.es

Jose Carlos Sancho Núñez
Universidad de Extremadura
Centro Universitario de Mérida
jcsanchon@unex.es

***Resumen*—Las amenazas en ciberseguridad siguen evolucionando, por lo que es esencial contar con plataformas que simulen ataques y vulnerabilidades, facilitando el aprendizaje continuo. Hack The Box es la plataforma líder en este ámbito, ofreciendo un entorno interactivo donde profesionales, investigadores y estudiantes pueden practicar y mejorar sus habilidades. Hack The Box se ha convertido en una comunidad global que fomenta el intercambio de conocimientos, la cooperación y la competencia sana. Estos elementos son cruciales para formar expertos en ciberseguridad. Este artículo presenta una experiencia real que demuestra cómo crear un espacio donde generar encuentros, no solo sirve como fuente de aprendizaje, sino como un motor de desarrollo colectivo en ciberseguridad, impulsado por su comunidad activa.**

***Index Terms*—Hack The Box, Comunidades, Capture The Flag, Ciberseguridad**

Tipo de contribución: *Formación e innovación educativa*

I. Introducción

En la actualidad, donde el número de ciberataques sigue creciendo día a día, la ciberseguridad es esencial para garantizar el funcionamiento de infraestructuras críticas y mantener la economía que se ha formado a través del ecosistema interconectado que ofrece la tecnología actual. El crimen organizado se ha hecho hueco en el ciberespacio, formando un sector profesionalizado y muy lucrativo que se estima que puede costarnos a los españoles unos 137.7 mil millones de dólares (USD) para el año 2028.[1]

La ciberseguridad es un sector en constante crecimiento en España, impulsado por la digitalización y el aumento de ciberataques. En 2023, se contrataron a más de 150.000 profesionales y alcanzó un valor de 2.000 millones de euros. A pesar de su crecimiento, la brecha de talento persiste, con más de 83.000 vacantes en 2024 y una proyección de 99.600 vacantes para 2025. La formación especializada es esencial para cubrir estas necesidades y asegurar el futuro del sector.[2] En el resto de Europa también persiste esta brecha: una reciente encuesta del *Eurobarometer* destaca que la brecha de talento en ciberseguridad en la Unión Europea es cada vez mayor, con un número significativo de organizaciones que luchan por contratar y retener personal especializado en ciberseguridad.

Para combatir esta brecha de talento bautizada como la *Cyber Skill Gap*[3], la Unión Europea (UE) ha apostado por la formación de nuevos profesionales a través de programas de enseñanza superior.[4] En marzo de 2025, la UE cuenta con 138 programas de educación entre todos los estados miembros repartidos en 32 programas de grado, 101 programas de máster y 5 programas posdoctorales. En España contamos con un programa de grado y 18 programas de máster, convirtiéndolo en el país con más programas de educación superior en ciberseguridad. [5]

Aparte de la formación en ciberseguridad impartida por las universidades, más de 120 centros educativos en España ofrecen titulaciones de Formación Profesional especializadas en ciberseguridad, tanto en la modalidad de Ciberseguridad en Entornos de las Tecnologías de la Información (CETI) como en Ciberseguridad en Entornos de las Tecnologías de la Operación (CETO). Estos programas permiten a los estudiantes adquirir habilidades prácticas y actualizadas en un sector con alta demanda de profesionales. [6]

Aunque la oferta de programas de formación en ciberseguridad es amplia, no se está consiguiendo que el número de alumnos matriculados aumente lo suficiente para poder cubrir las necesidades de este sector. Algunos de estos motivos podrían ser la falta de motivación e interés en estas carreras, ya sea por la barrera inicial que supone comenzar los estudios en ciberseguridad o la falta de interés frente a otros sectores más accesibles.

En el análisis de varios estudios realizado por J.B. Kim, Chen Zhong y Hong Liu [7], se lleva a cabo un experimento donde integran la gamificación en un curso de ciberseguridad de grado universitario con una duración de 4 meses. Comparando los resultados del aprendizaje entre los laboratorios gamificados y los tradicionales, así como las percepciones, la motivación y los intereses profesionales de los alumnos en experiencias de laboratorio de ciberseguridad gamificadas, descubren que los elementos competitivos de la gamificación parecen potenciar la motivación extrínseca que, según sus estudios, está vinculada a un mayor interés del sector entre los estudiantes.

Este artículo explora el impacto de *Hack The Box Meetups* como una comunidad estudiantil de jugadores de retos Capture

The Flag (CTF), que impulsa el aprendizaje y el desarrollo de habilidades en ciberseguridad. A través de la organización de eventos centrados en retos CTF, se ha logrado fomentar el interés en la ciberseguridad y motivar a los estudiantes a formar una carrera en este sector. Los eventos organizados por la comunidad han permitido a los participantes adquirir los conocimientos en ciberseguridad y las habilidades prácticas necesarias para realizar de forma autónoma retos CTF y comenzar a utilizar plataformas de formación en *Hacking* ético. Los resultados indican que la mayoría de los asistentes han encontrado las charlas accesibles e interesantes, además de valorar las actividades prácticas como entretenidas y de gran utilidad para aprender. Cabe destacar que algunos miembros de esta comunidad han empezado a participar en competiciones CTF y la mayoría de ellos han comenzado a utilizar plataformas de retos CTF después de participar en estos eventos. Estos resultados demuestran que la asistencia a estos eventos proporciona una experiencia valiosa para los alumnos de informática que enriquece su formación en ciberseguridad y les introduce al mundo del hacking ético y los retos CTF.

El trabajo se estructura de la siguiente forma: En la Sección II se describe el contexto y la motivación por la que se llevó a cabo este trabajo. En la Sección III se describe el programa de un evento, además del enfoque aplicado a las charlas y los recursos utilizados. En la Sección IV detallamos los eventos que se han celebrado con la comunidad. En la Sección V se presentan los resultados de estos eventos a través de los datos recopilados en los eventos y de una encuesta entregada a los participantes de este programa, junto con una discusión acerca de los datos obtenidos. Por último, en la Sección VI presentamos las conclusiones finales del artículo.

II. Motivación

El sector de la ciberseguridad ha sido históricamente un nicho aislado y de difícil acceso, especialmente en comparación con otros sectores más tradicionales, como la programación o el desarrollo. A menudo, se asocia la ciberseguridad con un grupo reducido de entusiastas que, en algunos casos, son percibidos como poco sociables o incluso malintencionados debido a la connotación errónea que se le atribuye al término *hackers*. Un ejemplo simple pero concreto de esto es la ausencia de la categoría ciberseguridad en plataformas populares como *Twitch*, a pesar de contar con una audiencia considerable. Sin embargo, en otras plataformas con un enfoque más alternativo, como *Kick*, esta categoría sí ha sido creada.

Este fenómeno también se refleja a nivel local, por ejemplo, en Cáceres, donde no existen grupos abiertos en ciberseguridad que perduren en el tiempo, permitiendo formar parte de una comunidad activa. La falta de tales grupos limita la posibilidad de compartir conocimientos, experiencias, aprender, enseñar, realizar networking, generar vínculos personales y, lo más importante, mantener estándares éticos, una característica esencial para este rubro.

A nivel regional, se ha observado que la Universidad de Extremadura no cuenta con una comunidad de ciberseguridad que participe activamente en competiciones grupales, como el caso de los organizados por las Jornadas Nacionales de Investigación en Ciberseguridad (JNIC), las cuales se celebran anualmente y cuentan con la participación de universidades tanto nacionales como internacionales, tal como se muestra en la Tabla I. Estas competiciones son un referente en el ámbito de la ciberseguridad y representan una excelente oportunidad para que las instituciones educativas demuestren la calidad de sus programas formativos y la preparación de sus estudiantes en este campo. La falta de presencia destacada de la Universidad de Extremadura en tales eventos refleja la necesidad de fortalecer la formación y el apoyo a los estudiantes en la región, con el fin de mejorar su competitividad en el ámbito de la ciberseguridad, tanto a nivel nacional como internacional.

Tabla I
Participación CTF ediciones JNIC

Edicion	Modalidad	Alcance	Participantes	Equipos
2024	Equipos	Internacional	326	181
2023	Equipos	Internacional	272	169
2022	Equipos	Internacional	277	93
2021	Individual	Internacional	214	NO
2019	Individual	Nacional	97	NO

Ante esta situación, se identificó la necesidad de crear un espacio que permita la formación de un grupo sostenible a lo largo del tiempo, ya que muchas veces, debido a la falta de apoyo económico y de un equipo comprometido, estos grupos tienden a desaparecer rápidamente. El objetivo es construir una comunidad inclusiva en la que cualquier persona, sin importar su nivel de experiencia previa, pueda participar en actividades relacionadas con la ciberseguridad. De esta forma, se busca impulsar el potencial de talentos en ciberseguridad en la región, guiando a los miembros en sus primeros pasos en la disciplina, sin la intención de obtener retribuciones económicas, promoviendo un ambiente más colaborativo y fraternal, alejado de la estructura tradicional de alumno-profesor.

Uno de los principales nichos para reclutar miembros son las universidades y los centros de formación profesional, especialmente en las áreas tecnológicas, como las carreras de ingeniería informática, telecomunicaciones e inteligencia artificial. Estos estudiantes suelen tener poca o ninguna experiencia previa en ciberseguridad, lo que representa una oportunidad clave para despertar su interés en esta disciplina. Sin embargo, carecen de una comunidad en la que puedan buscar compañeros para formar equipos, compartir conocimientos y experiencias, lo que limita su desarrollo en este campo. La falta de apoyo a los principiantes constituye una barrera significativa, siendo una de las principales razones del abandono de esta disciplina.

Con el fin de abordar esta necesidad, se presentó la oportunidad de formar parte de la comunidad Hack The Box Meetup, a través de un miembro de la comunidad Hack The Box Meetup: Salta, AR [8]. Los Grupos de Meetup de HTB son una red global de comunidades dedicadas al hacking ético, establecidas en ciudades de todo el mundo, que se reúnen mensualmente para realizar talleres, charlas y competiciones de CTF (Capture The Flag). Los miembros de esta comunidad tienen la oportunidad de conectarse, practicar hacking y aprender [9].

II-A. Hack The Box

Hack The Box (HTB) es una plataforma que ofrece a sus usuarios una amplia variedad de retos en ciberseguridad, que

abarcan diferentes categorías, como explotación de vulnerabilidades web, ingeniería inversa, criptografía, entre otras. A través de máquinas virtuales diseñadas específicamente y retos que varían en niveles de dificultad, desde principiante hasta experto, HTB permite a los usuarios realizar pruebas CTF en un entorno seguro y controlado, donde pueden aplicar y perfeccionar sus habilidades sin comprometer la seguridad de sistemas reales.

HTB brinda a los participantes la oportunidad de aprender de sus errores, ya que no solo pueden repetir los retos, sino que también tienen acceso a pistas y la posibilidad de investigar soluciones compartidas por la comunidad. Este proceso de prueba y error constituye una de las bases fundamentales del aprendizaje en ciberseguridad, permitiendo a los usuarios perfeccionar sus habilidades de manera más efectiva en comparación con otros entornos de aprendizaje tradicionales.

II-B. El éxito de la Gamificacion

Una de las cosas más interesantes de HTB es la gamificación, que es una estrategia educativa que utiliza elementos y dinámicas de juego para mejorar el aprendizaje y la motivación en el ámbito de la ciberseguridad.

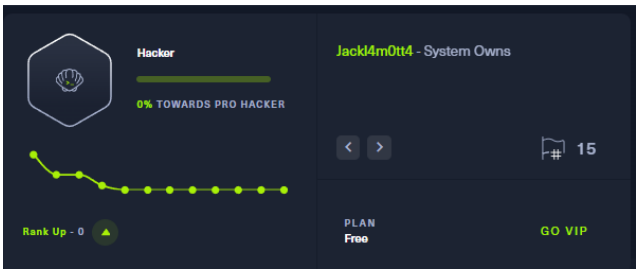


Figura 1. Sistema de Rango

Una de las principales ventajas de la gamificación en *Hack The Box* (HTB) es su capacidad para motivar y mantener la constancia en la resolución de las máquinas vulnerables y retos. A través de la implementación de sistemas de puntuación y recompensas, como se observa en la Figura 1 los usuarios se sienten impulsados a resolver más desafíos para mejorar su desempeño, lo que incentiva la práctica continua. La competencia y las clasificaciones ofrecen un marco en el cual los participantes pueden medir su progreso y compararse con otros, lo que fomenta un sentido de logro y mejora constante a través de la competitividad.

Esto permite crear un *ranking* no solo a nivel individual, sino también a nivel grupal, en forma de equipos, universidades y países.

II-C. Hack The Box España

España cuenta con 2022 miembros en la plataforma Hack The Box (HTB), quienes han vulnerado un total de 69,607 máquinas y resuelto 24,679 retos. En la Figura 2 se presentan los tres primeros usuarios actuales que lideran la tabla.

A nivel de universidades, en la Figura 3 se observan las universidades que forman parte del top 5.

Como se puede observar, la Universidad de Valladolid ocupa el quinto lugar en el ranking, compuesto actualmente por 15 miembros.

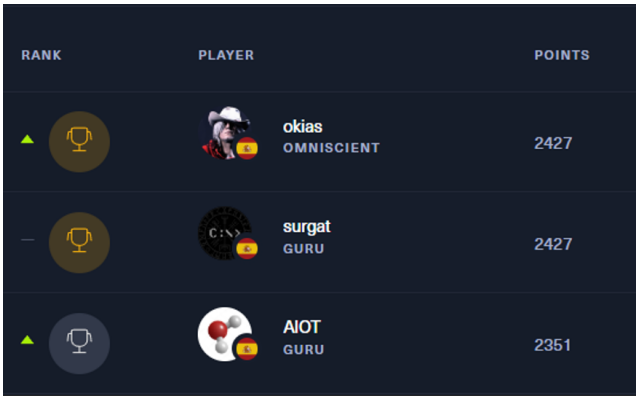


Figura 2. Top 3 HTB en España

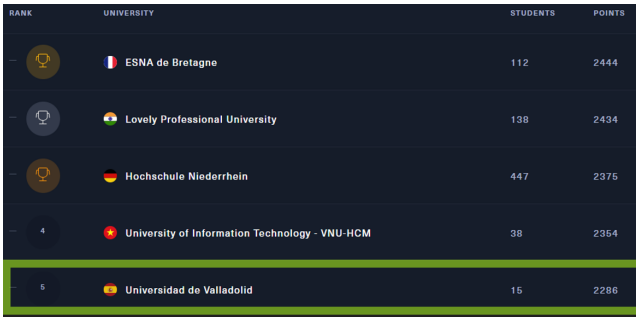


Figura 3. Ranking Universidades

II-D. Comunidad de HTB

No solo su sistema de gamificación y la riqueza del contenido técnico destacan a *Hack The Box* como una de las plataformas pioneras en el aprendizaje de ciberseguridad, sino que lo que realmente distingue a HTB de otras plataformas es el enfoque en la comunidad que logra crear entre sus miembros. Mientras que la mayoría de las plataformas de *hacking* se enfocan principalmente en el reto individual, *HTB* fomenta una interacción constante entre los usuarios, facilitando el intercambio de ideas, consejos y soluciones a través de foros, canales de *Discord* y otros espacios colaborativos. Este enfoque convierte a la plataforma en un punto de encuentro donde los conocimientos no solo se comparten con el fin de resolver retos, sino también en el proceso de aprendizaje y crecimiento colectivo.

HTB fomenta la competencia sana, genera espacios seguros e inclusivos para compartir experiencias y conocimientos, y ofrece oportunidades para formar grupos para realizar *networking* o participar en actividades fuera del ámbito del *hacking*. Estas características aportan numerosas ventajas, tanto para los profesionales del sector como para los aprendices y las organizaciones.

II-E. HTB Meetup por el mundo

Los grupos de *Hack The Box Meetups* son de carácter global y constituyen una de las mejores maneras de conectarse con la comunidad. Actualmente, la plataforma cuenta con casi 34,000 miembros, distribuidos en 140 grupos ubicados en 69 países [9].

Cada mes, se organizan múltiples eventos en todo el mundo



Figura 4. Grupos de HTB a nivel mundial

con la colaboración de los embajadores de reuniones. El objetivo principal de estos eventos es facilitar la conexión de los usuarios con un grupo local al que puedan pertenecer, un grupo que les ayudará a aprender, mejorar sus habilidades de manera práctica y brindarse apoyo e inspiración mutua.

Estos eventos gratuitos tienen un valor real y práctico, con el propósito de que los participantes salgan de cada encuentro habiendo aprendido algo nuevo.

II-F. HTB Meetup en España

Aunque la plataforma de *Hack The Box* (HTB) está disponible principalmente en inglés, cuenta con una cantidad considerable de miembros provenientes de España y América Latina, lo que demuestra que el idioma no representa una barrera para la ciberseguridad.

Un aspecto destacable es que la Cátedra *INCIBE-UEX-EPCC* y los alumnos de Doctorado jugaron un papel crucial en la creación del primer grupo de *meetups* de HTB en España, en la Universidad de Extremadura.

A partir de este grupo, se colaboró con *Hack The Box* para abrir más grupos de *meetups* a lo largo de España, como se observa en la Figura 5. No sería extraño que en un futuro cercano continúen apareciendo nuevos grupos en diferentes regiones.

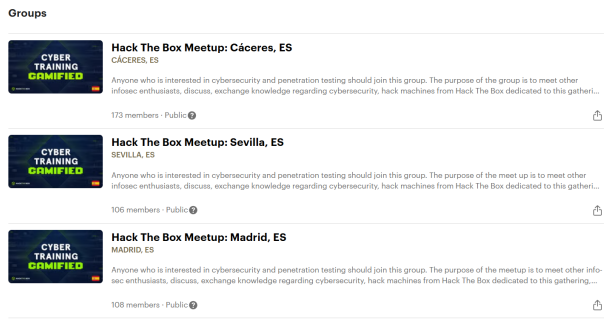


Figura 5. HTB Meetups España

II-G. Catredra INCIBE-UEX-EPCC

La Cátedra tiene como misión promover el conocimiento y la cultura de la ciberseguridad en Extremadura, mediante la realización de actividades formativas y de investigación en esta área [10]. Asimismo, busca proteger a las empresas, organismos y ciudadanos de la región de Extremadura frente a las amenazas emergentes del mundo digital.

A partir de esta colaboración, se ha hecho posible la creación del grupo de Hack The Box en Cáceres, lo que permite proporcionar el espacio y los recursos necesarios para llevar a cabo dicha misión de manera mensual, organizar eventos y fomentar el aprendizaje en comunidad.

Es importante destacar que, sin un espacio físico adecuado, los recursos necesarios de conectividad para llevar a cabo los eventos y un personal dispuesto a organizar las actividades sin retribución económica, se dificultaría la continuidad y sostenibilidad del grupo. Estos factores son claves para mantener el grupo activo a lo largo del tiempo, lo cual representa uno de los mayores desafíos.

III. Descripción de los eventos

III-A. Enfoque pedagógico

El objetivo principal de estos eventos es presentar a los participantes conceptos, técnicas y áreas de estudio en ciberseguridad de forma práctica, con las que luego pueden aplicar en plataformas de aprendizaje gamificadas como Hack The Box o en eventos CTF. De esta forma, conseguimos que el alumnado supere la barrera inicial que supone realizar por primera vez retos CTF e incentivarlos a que participen en competiciones (individuales o grupales), aprovechando el poder de difusión de estos espacios y la formación de una comunidad con integrantes interesados en participar en competiciones relacionadas con la ciberseguridad.

III-B. Estructura de la actividad

Estos eventos, organizados mensualmente de forma presencial y remota, suelen tener una duración de 2 horas. En ellos, normalmente se cuenta con un instructor principal que es quien conduce el evento y otros instructores de apoyo para ayudar a los participantes en las tareas prácticas. Estas actividades solo se enfocan en una temática con la que se trabajará en todo el evento y se pueden dividir en tres secciones: una introducción teórica, una práctica guiada y una pequeña competición entre los participantes.

La introducción teórica es impartida por el instructor principal, presentando los conceptos básicos del tema y la información necesaria para poder comprender luego la actividad práctica. Se intenta que no tenga una duración muy extensa, ya que los participantes pueden tener los conocimientos teóricos de antemano, pero carezcan de experiencia práctica y de cómo se ejecutan estos conceptos en la realidad.

Una vez finalizada la introducción teórica, los participantes realizarán una prueba CTF guiada por el instructor principal. El instructor, además de presentar los objetivos finales de la prueba, también indicará metas parciales que guiarán a los alumnos sobre cómo se debe realizar la prueba paso a paso. De esta forma, el flujo de trabajo de esta práctica guiada es el siguiente: el instructor principal presentará la meta parcial y la completará ante los participantes como demostración. Después, los alumnos lo realizarán en sus equipos, dándoles tiempo para que lo realicen, pedir ayuda a los instructores de apoyo y resolver dudas con la ayuda de los instructores. Una vez pasado el tiempo, se pasa a la siguiente meta y así hasta que los alumnos completen el reto.

Por último y para finalizar el evento, se les presenta a los alumnos un reto CTF que tienen que realizar de forma autónoma, con una recompensa para los primeros participantes que lo completen a forma de simulacro de CTF. El añadido del aliciente ayuda a que los participantes completen el reto con una restricción de tiempo autoimpuesta a forma de ensayo

Tabla II
EVENTOS CELEBRADOS

#	Evento	Modalidad	Asistentes totales	Asistentes en presencial	Asistentes en remoto	Tema principal
1	Hack The Box Meetup Cáceres #1	Charla presencial	64	64	-	Metodología pentesting
2	Hack The Box Meetup #2: Hacking Web part. 1	Charla Híbrida	59	32	27	Web Hacking
3	Hack The Box Meetup #3: Explorando el Universo CTF	Charla presencial	26	26	-	Categorías CTF
4	HackTheBox Meetup Cáceres #4: Entrypoint León, ES.	Charla híbrida + CTF Presencial	36	23	13	Investigación + CTF
5	Hack The Box Meetup #5: Hacking Web con BurpSuite	Charla Híbrida	26	22	4	Web Hacking

ante competiciones CTF, impulsando a los participantes a que utilicen estrategias tanto a nivel individual como de forma grupal si el reto se realiza por grupos. El premio en estos casos suele ser productos cedidos por Hack The Box u otro material que también ayude al ganador a seguir formándose en ciberseguridad como libros u otro material.

III-C. Recursos utilizados

Gracias a que esta iniciativa es organizada por la plataforma Hack The Box, ésta nos proporciona la infraestructura necesaria para realizar las actividades, material formativo y premios que pueden recibir tanto los participantes como los organizadores de estas comunidades.

Para la realización de las pruebas prácticas y las competiciones CTF, se utiliza la plataforma Enterprise de Hack The Box, [11] la cual nos permite que los alumnos puedan realizar las prácticas desde una máquina virtual desplegada en su equipo o incluso desde un navegador si el alumnado no pudiera desplegar la máquina virtual, ya sea por falta de recursos del equipo o que les suponga una barrera técnica.

Para los instructores, la plataforma cuenta con una multitud de pruebas CTF de diferentes temáticas y dificultades, lo que se traduce en un ahorro de tiempo en la preparación del material al no tener que implementar los retos. También ofrece logs y clasificaciones de la actividad de los participantes para poder realizar un seguimiento y alentar a los alumnos a la competitividad.

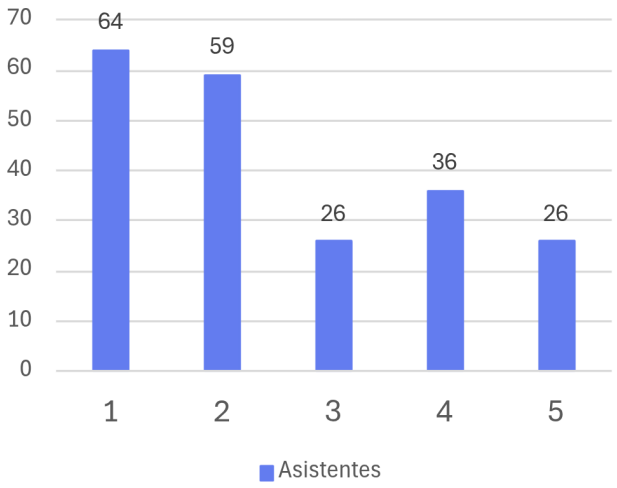


Figura 6. Asistentes por evento

IV. EVENTOS COMPLETADOS

La Tabla II presenta los eventos realizados por la comunidad Hack The Box Meetup Cáceres en orden cronológico. En total se han celebrado 5 eventos: 4 en la ciudad de Cáceres y 1 en León. Como se ha mencionado anteriormente, los eventos de *Hack The Box Meetups* se organizan de forma mensual en modalidad híbrida, siendo posible asistir tanto de forma presencial como en remoto. Esto último depende de la disposición del equipo necesario para retransmitirlo. En nuestro caso se han celebrado 3 charlas en modalidad híbrida, en los eventos #2, #4 y #5.

En los cinco eventos organizados por *Hack The Box Meetups*, se ha observado en los asistentes una participación activa e interés por los temas tratados tanto de forma presencial como remota. La modalidad híbrida ha sido fundamental para mejorar la accesibilidad y la participación de los eventos, posibilitando a los alumnos participar en las charlas y actividades en caso de no poder asistir de forma presencial.

Los temas abordados en los eventos han sido variados, desde la introducción al *hacking* ético siguiendo metodologías aplicadas al pentesting hasta las distintas técnicas de web *hacking*, pasando por la resolución de diferentes categorías comunes en los eventos de CTF. El objetivo de esta variedad de temas es proporcionar a los participantes una amplia gama de conocimientos y habilidades, adaptándose a distintos niveles de experiencia e intereses.

El primer evento, centrado en la metodología de pentesting, tuvo una asistencia completa de 64 participantes en presencial. Este evento se centró en introducir a los participantes en las técnicas y procesos utilizados en *pentesting* para poder aplicarlos en retos CTF del estilo *Full Pwn*.

El segundo evento, que abordó la primera parte de una serie de eventos sobre *web hacking*, asistieron un total de 59 participantes, de los cuales 32 asistieron de forma presencial y 27 de forma remota.

El tercer evento, dedicado a las categorías de retos Capture The Flag (CTF), tuvo una asistencia de 26 participantes en presencial. Este evento ofreció una visión general de los tipos de desafíos que pueden encontrarse en competiciones de CTF.

El cuarto evento, celebrado en León, combinó una charla híbrida con una competición CTF en presencial. Asistieron en total 36 participantes, 23 de forma presencial y 13 de forma remota. Este evento combinó una charla sobre los progresos en la investigación en juguetes conectados y entornos 5G con un CTF de modalidad *Full Pwn*.

El quinto evento, centrado en el uso de BurpSuite para el hacking web, contó con 26 participantes, 22 de forma presencial y 4 de forma remota. Este evento sirvió como

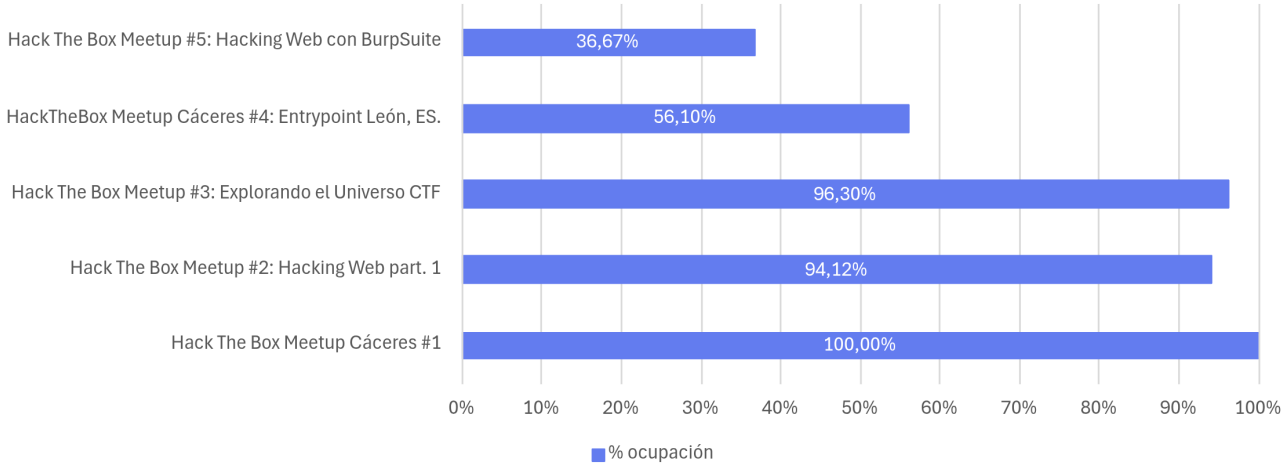


Figura 7. Porcentaje de aforo completado por evento

introducción al uso de la herramienta BurpSuite para utilizarla en *web hacking*.

La tabla III recoge los datos de participación en las competiciones CTF de los miembros de la comunidad *Hack The Box Meetup* Cáceres. Aunque en la tabla se presentan los datos de rendimiento del equipo en las competiciones, lo más destacado es la participación de los alumnos en este tipo de competiciones gracias a los eventos organizados por la comunidad. Gracias a la comunidad formada por los alumnos, estos han podido encontrar a compañeros con los que pueden formar equipos y colaborar entre todos para participar en competiciones CTF.

Tabla III
PARTICIPACIÓN EN COMPETICIONES CTF

CTF	Participantes	Banderas logradas	Posición
HTB University CTF 2025	14	30/49	148/872
HTB Cyber Apocalypse CTF 2025	10	29/77	546/8129

V. RESULTADOS

En base a los datos recogidos en los eventos a través de la plataforma *Meetup*, se muestra el número de participantes por cada evento en la Figura 6, asociando cada evento por el número identificativo presente en la tabla II. Aquí se contabiliza como asistentes tanto los registrados en la plataforma *meetup* que han atendido de forma presencial como los usuarios que han atendido de forma remota. Aunque la tendencia parece que sea a la baja, alguno de los factores que limitan la asistencia de los eventos es el aforo disponible de la sala donde se impartía cada evento, la difusión con la que contábamos en cada evento y la posibilidad de atender al evento de forma remota. En la Figura 7 se muestra la ocupación del aforo de la ubicación donde se realizaron los *meetups*. Se puede observar que los 3 primeros eventos contaron con una asistencia casi completa. El evento número 4, al realizarse en otra ciudad, no se contaba con la suficiente difusión. Para este último evento no se ha podido identificar la diferencia de aforo completado con respecto a *meetups* anteriores.

Para entender la experiencia de los alumnos y estudiar la efectividad de estos eventos, se ha difundido una encuesta a los integrantes del grupo *Hack The Box Meetup: Cáceres, ES* registrados en *meetup.com* [12] que hayan asistido al menos a uno de los eventos organizados por el grupo. La encuesta consiste en componentes cuantitativos con preguntas basadas en la escala de Likert y de selección, tanto única como múltiple. En total, hemos recibido 17 respuestas de los participantes. En esta sección se analizarán las respuestas de la encuesta en busca de resultados e ideas que puedan beneficiar a este proyecto.

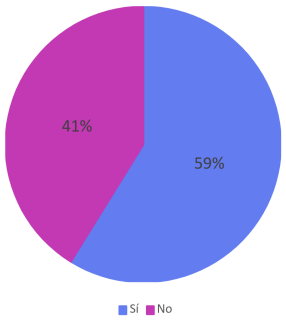


Figura 8. Respuestas a la pregunta: "¿Tenía alguna experiencia sobre *hacking* ético y/o retos *Capture The Flag* (CTF) antes de participar en los *meetups* de Hack The Box Cáceres?"

Como se muestra en la Figura 8, hay un gran sector de los participantes que no han tenido experiencia previa sobre *hacking* ético y/o retos *Capture The Flag* (CTF) y la mayoría de ellos está cursando un grado universitario según la Figura 9.

La Figura 10 expone las respuestas de la conformidad de los enunciados según la escala de Likert. A partir de esta figura, podemos destacar los siguientes resultados:

1. Los participantes han podido seguir la charla con facilidad, comprender el temario y aprender cosas que no conocían según los enunciados 1, 2, 3, 4, 6, 7, 8 y 9.
2. Las pruebas CTF realizadas estaban ajustadas a su dificultad y les ha sido entretenido según los enunciados 5, 11, 12, 13 y 14.

3. En el enunciado 10, el 53,3 % de los encuestados han contestado que están de acuerdo o muy de acuerdo al enunciado ”Me gustaría dedicarme profesionalmente a la ciberseguridadz un 41,2 % han seleccionado la opción ”neutral”. Aunque la mayoría de los encuestados quieren dedicarse profesionalmente a la ciberseguridad, hay un gran número de alumnos que no se han decidido aún, y estas charlas podrían ser el instrumento para que decidan su futuro.
4. Como figura en el enunciado 15, el 64,7 % de los participantes han empezado a utilizar plataformas de retos CTF despúes de participar en los *meetups*, con lo que estas charlas les ha proporcionado las herramientas y la motivación suficientes para que los alumnos les dedique su tiempo libre para formarse en ciberseguridad y ganar experiencia práctica a través de plataformas de retos CTF.

Cabe destacar que varios de los asistentes,luego de asistir a varios eventos, se animaron a preparar temas para compartir con el grupo, permitiendo seguir creciendo en comunidad. Además, algunos integrantes de la comunidad han participado en dos competiciones CTF organizadas por Hack The Box como un grupo.

Por último, la valoración media de la experiencia de estos eventos según los encuestados es de un 4.47 sobre 5. Con estos resultados podemos concluir que la asistencia a este evento proporciona a los participantes una experiencia valiosa que les puede fortalecer su aprendizaje en ciberseguridad, además de introducirles al mundo de los retos CTF.

VI. Conclusiones

Los resultados indican que la mayoría de los asistentes han encontrado las charlas útiles y valiosas para su formación en el campo de la ciberseguridad. También, los resultados de la encuesta reflejan un aumento de la implicación de los alumnos en actividades extracurriculares sobre ciberseguridad y hacking ético, como el uso de plataformas de formación gamificadas y la participación en competiciones *Capture The Flag*. Aunque un gran número de estudiantes aún no han decidido si labrar su futuro profesional en el sector de la ciberseguridad, la participación en actividades programadas por esta comunidad puede ayudar a que tomen esta decisión.

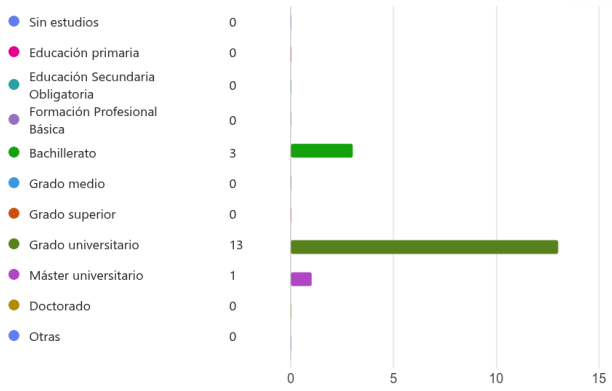


Figura 9. Respuestas a la pregunta: ”¿Cuál es su nivel de estudios? Elija el nivel que está cursando actualmente”



Figura 10. Respuestas a la pregunta: ”Tras participar en alguno de los meetups de Hack The Box Meetup Cáceres, ¿en qué medida está de acuerdo con estas afirmaciones?”

Este proyecto sigue en desarrollo con el objetivo de tratar más temas relacionados con la ciberseguridad, explorando temas de operaciones de seguridad informática defensiva y de expandir este tipo de comunidades a otras localidades de la región y ayudar a la formación de nuevas comunidades en otras regiones del territorio nacional español.

Agradecimientos

Esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation) - Instituto Nacional de Ciberseguridad (INCIBE) en el proyecto Proyecto C109/23 ”Proyecto estratégico UEx (Escuela Politécnica de Cáceres)-INCIBE”.

Referencias

[1] Stadista, “Spain cost of cybercrime 2028 — statistaestimated annual cost of cybercrime in spain from 2016 to 2028 (in billion u.s. dollars) [graph],” 9 2023. [Online]. Available: https://www.statista.com/forecasts/1429612/spain-cybercrime-cost-annual

[2] C. A. Brugal, “Análisis y diagnóstico del talento de ciberseguridad en españa,” Instituto Nacional de Ciberseguridad, Tech. Rep., 2022. [Online]. Available: https://files.incibe.es/incibe/talento/INCIBE_InformeCompleto_DIAG.pdf

[3] C. A. Ramezan, “Examining the cyber skills gap: An analysis of cybersecurity positions by sub-field,” *Journal of Information Systems Education*, vol. 34, no. 1, pp. 94–105, 2023. [Online]. Available: Availableat:https://aisel.aisnet.org/jise/vol34/iss1/8

[4] European Union Agency for Cybersecurity, J. R. Nurse, K. Adamos, A. Grammatopoulos, and F. Di Franco, *Addressing the EU cybersecurity skills shortage and gap through higher education*, 2021.

[5] ENISA, “Cyberhead programmes evolution and statistics.” [Online]. Available: https://tools.enisa.europa.eu/topics/education/cyberhead/#/statistics

[6] INCIBE, “Instituciones que imparten formación en ciberseguridad en españa,” 2024. [Online]. Available: https://www.incibe.es/sites/default/files/paginas/talento/catalogos-formacion/catalogo_instituciones_24.pdf

[7] J. J. B. Kim, C. Zhong, and H. Liu, “The impact of gamification on cybersecurity learning: Multi-study analysis,” *Communications of the Association for Information Systems*, vol. 56, pp. 57–96, 2025. [Online]. Available: https://aisel.aisnet.org/cais/vol56/iss1/6/

[8] Meetup, “Hack the box meetup: Salta, ar — meetup.” [Online]. Available: https://www.meetup.com/hack-the-box-meetup-salta-ar/

[9] Hack The Box, "Hack the box community meetups — let's hacking & train together." [Online]. Available: https://www.hackthebox.com/meetups

[10] A. C. Lindo, "Una apuesta por la educación en ciberseguridad desde el ámbito universitario," in *Actas de las primeras Jornadas Nacionales de Investigación en Ciberseguridad: León, 14, 15, 16 de septiembre de 2015. I JNIC2015, 2015, ISBN 978-84-9773-742-5, págs. 198-202*, 2015, pp. 198–202. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=5469843

[11] Hack The Box, "Cyber labs & cyber exercises — real-world simulated training." [Online]. Available: https://www.hackthebox.com/business/professional-labs

[12] Meetup, "Hack the box meetup: Cáceres, es — meetup." [Online]. Available: https://www.meetup.com/hack-the-box-meetup-caceres-es/

Estrategias de securización en el Cloud

Badreddine Abbaoui Toufiq
Universidad de Extremadura
Escuela Politecnica de Cáceres
badreddine@unex.es

Mohammadhossein Homaei
Universidad de Extremadura
Escuela Politecnica de Cáceres
mhomaein@alumnos.unex.es

Marcos Jesús Sequera Fernández
Universidad de Extremadura
Escuela Politecnica de Cáceres
marcosjesus@unex.es

María Mar Ávila Vegas
Universidad de Extremadura
Escuela Politecnica de Cáceres
mmavila@unex.es

***Resumen*—La adopción masiva de servicios en la nube ha supuesto una transformación radical en la gestión de infraestructuras tecnológicas, ofreciendo ventajas significativas en términos de escalabilidad, disponibilidad y eficiencia. No obstante, esta evolución también ha introducido nuevos desafíos relacionados con la ciberseguridad, especialmente en entornos donde los datos sensibles y las operaciones críticas se trasladan a plataformas compartidas.**

Este trabajo presenta el diseño e implementación de una arquitectura segura en Amazon Web Services (AWS) utilizando la metodología de Infraestructura como Código (IaC) con Terraform. La propuesta contempla la contenerización de una aplicación web desarrollada en Flask, el despliegue de instancias EC2 mediante AMIs personalizadas, la configuración de una red privada segmentada mediante VPCs y subredes, y la integración de mecanismos de protección como AWS WAF y balanceadores de carga.

El resultado es una infraestructura modular, automatizada y replicable, orientada a entornos de formación y prácticas de ciberseguridad, que garantiza la protección de los servicios desplegados frente a accesos no autorizados y amenazas comunes. Además, se validan todas las configuraciones mediante pruebas de conectividad, acceso y resistencia ante ataques, demostrando la viabilidad del enfoque adoptado y sentando las bases para futuras ampliaciones y optimizaciones.

***Index Terms*—Ciberseguridad, Cloud Computing, AWS, Terraform, Infraestructura como Código.**

Tipo de contribución: *Formación e innovación educativa*

I. Introducción

El avance del *Cloud Computing* ha transformado radicalmente la gestión de infraestructuras tecnológicas en organizaciones de todos los tamaños, facilitando el acceso a recursos computacionales escalables y bajo demanda. Esta adopción masiva ha permitido optimizar costes, mejorar la flexibilidad operativa y acelerar los ciclos de desarrollo de software. Sin embargo, la naturaleza dinámica y compartida de los entornos cloud también ha introducido nuevos vectores de ataque y desafíos en materia de ciberseguridad.

En este contexto, garantizar la seguridad de los sistemas desplegados en la nube se ha convertido en una prioridad crítica. Las amenazas a la confidencialidad, integridad y disponibilidad de los datos exigen arquitecturas diseñadas con principios de seguridad desde el origen. Para abordar esta problemática, el uso de técnicas de IaC se ha consolidado como una estrategia eficaz para desplegar entornos seguros, replicables y auditables.

Este trabajo presenta el diseño e implementación de una infraestructura segura sobre Amazon Web Services, automatizada mediante Terraform y orientada a la ejecución de una aplicación web contenerizada. La propuesta incluye mecanismos de segmentación de red, control de acceso y protección frente a amenazas web mediante AWS WAF.

El objetivo principal del trabajo es demostrar cómo una arquitectura basada en buenas prácticas de securización cloud y aprovisionamiento automático puede utilizarse eficazmente en contextos formativos, permitiendo a los estudiantes experimentar en entornos realistas y personalizados. En particular, la solución propuesta facilita la creación de escenarios individualizados para asignaturas relacionadas con redes de comunicaciones y seguridad, ofreciendo una herramienta docente práctica para evaluar competencias específicas del alumnado.

La contribución distintiva de este trabajo frente a otros existentes radica en:

- El enfoque claramente orientado a la aplicación docente y práctica de la arquitectura propuesta.
- La automatización integral mediante Terraform, que garantiza replicabilidad, consistencia y escalabilidad.
- La validación empírica del sistema en un entorno formativo real, analizando su utilidad como herramienta educativa.

El resto del documento se estructura de la siguiente manera: en la Sección II se presenta el estado del arte relacionado con tecnologías clave y trabajos previos en IaC y seguridad cloud; la Sección III detalla los materiales y métodos empleados; la Sección IV describe el diseño e implementación técnica de la solución propuesta; la Sección V muestra los resultados obtenidos en la validación práctica y su potencial aplicación docente; finalmente, en la Sección VI se presentan las conclusiones obtenidas y líneas futuras de investigación y mejora.

II. Estado del Arte

Este apartado presenta una revisión de las tecnologías y herramientas más relevantes para la implementación de infraestructuras seguras en entornos de computación en la nube. El análisis incluye la evaluación de proveedores de servicios Cloud, herramientas de Infraestructura como Código y frameworks de desarrollo web, con el objetivo de identificar las soluciones más adecuadas para desplegar arquitecturas seguras, automatizadas y escalables.

Antes de realizar la comparativa, es importante destacar los beneficios clave que la Infraestructura como Código aporta en el contexto de la seguridad en la nube:

- **Automatización y consistencia**: Reduce la intervención manual y los errores humanos, asegurando despliegues uniformes y controlados.
- **Gestión de acceso y redes seguras**: Facilita la creación de VPCs, subredes, grupos de seguridad y otras configuraciones de red de forma programática.
- **Escalabilidad y recuperación ante desastres**: Permite la réplica rápida de infraestructuras y la implementación de estrategias de respaldo.
- **Integración con servicios de seguridad**: Compatible con servicios como AWS WAF, IAM o herramientas de monitorización para fortalecer las defensas del entorno cloud.

II-A. Comparación de Proveedores de Servicios en la Nube

Los principales proveedores de servicios cloud presentan diferencias clave en términos de cuota de mercado, soporte empresarial, nivel de seguridad y compatibilidad con herramientas IaC. La Tabla I muestra una comparativa entre Amazon Web Services, Microsoft Azure y Google Cloud Platform (GCP), destacando que AWS lidera el mercado con un amplio ecosistema y fuerte integración con herramientas IaC.

Tabla I
COMPARACIÓN DE PROVEEDORES CLOUD

Característica	AWS	Azure	GCP
Cuota de Mercado	32 %	22 %	11 %
Integración con IaC	Alta	Media	Alta
Soporte Empresarial	Excelente	Muy bueno	Bueno
Servicios Gestionados	Amplia variedad	Buena variedad	Centrado en IA
Seguridad	Alta	Alta	Alta

II-B. Comparación de Herramientas de IaC

La automatización del despliegue y la gestión de infraestructuras se ha consolidado como una necesidad crítica en entornos cloud. En la Tabla II se comparan las herramientas más populares de IaC: Terraform, Ansible, Puppet y Chef. Terraform destaca por su lenguaje declarativo (HCL), su alta escalabilidad y una comunidad muy activa, lo que lo convierte en una elección sólida para la gestión de infraestructuras complejas en la nube.

Tabla II
COMPARACIÓN DE HERRAMIENTAS DE IAC

Característica	Terraform	Ansible	Puppet	Chef
Tipo	Declarativa	Imperativa	Declarativa	Imperativa
Lenguaje	HCL	YAML	DSL	Ruby
Escalabilidad	Alta	Media	Alta	Alta
Comunidad	Grande	Grande	Media	Media

II-C. Comparación de Frameworks de Desarrollo Web

En el contexto del desarrollo de aplicaciones web seguras y escalables, la elección del framework es un factor clave. La Tabla III presenta una comparación entre Flask, Express.js, Django y Ruby on Rails. Flask ha sido seleccionado para este proyecto debido a su simplicidad, flexibilidad y facilidad de integración con contenedores Docker y entornos de infraestructura automatizada.

Tabla III
COMPARACIÓN DE FRAMEWORKS WEB

Característica	Flask	Express.js	Django	Ruby on Rails
Lenguaje	Python	JavaScript	Python	Ruby
Simplicidad	Alta	Alta	Media	Media
Escalabilidad	Media	Alta	Alta	Alta
Facilidad de Uso	Alta	Alta	Media	Baja

II-D. Selección de Tecnologías

A partir del análisis anterior, se concluye que AWS, Terraform y Flask constituyen la combinación más adecuada para este proyecto, dado su equilibrio entre seguridad, automatización, escalabilidad y facilidad de uso. Esta selección permite diseñar una arquitectura robusta y fácilmente mantenible, facilitando la incorporación de mecanismos de protección como WAF, balanceadores de carga y subredes privadas desde el inicio del despliegue.

III. MATERIALES

Para la implementación de una infraestructura segura en la nube, se han empleado diversas herramientas y tecnologías que permiten la gestión automatizada y la protección de los recursos desplegados. A continuación, se detallan los materiales utilizados:

III-1. Plataforma Cloud: Amazon Web Services: Como se menciono en la seccion II, se ha seleccionado Amazon Web Services como la plataforma de despliegue debido a su escalabilidad, seguridad y diversidad de servicios. Los servicios utilizados incluyen:

- **Elastic Compute Cloud (EC2)**[1]: Servicio de máquinas virtuales utilizado para desplegar tanto la aplicación web como la base de datos. Las instancias EC2 forman el núcleo del sistema computacional en la nube.
- **Virtual Private Cloud (VPC)**[2]: Proporciona una red virtual aislada dentro de AWS. Permite definir subredes públicas y privadas, establecer tablas de enrutamiento, gateways y gestionar el flujo de tráfico de red de forma segura.
- **Security Groups**[3]: Actúan como firewalls virtuales que controlan el tráfico entrante y saliente de las instancias EC2. Se configuran reglas específicas para restringir el acceso únicamente a direcciones IP, puertos o servicios autorizados, lo que refuerza la postura de seguridad de la infraestructura.
- **Application Load Balancer (ALB)**[4]: Se encarga de distribuir automáticamente el tráfico de la aplicación entre múltiples instancias EC2, mejorando la disponibilidad y escalabilidad del sistema. También facilita la integración con servicios como WAF.
- **Target Group**[5]: Conjunto lógico de instancias EC2 que reciben tráfico desde el ALB. Cada target group define el protocolo, el puerto y las condiciones de verificación de salud, asegurando que solo las instancias operativas reciban tráfico.

- **AWS Web Application Firewall (WAF)**[6]: Proporciona protección a nivel de capa 7, filtrando el tráfico HTTP y bloqueando ataques web comunes como inyecciones SQL, ataques XSS o accesos no autorizados. Se integra directamente con el ALB.

III-A. Infraestructura como Código

Se ha utilizado Terraform para la automatización del despliegue, asegurando que la infraestructura sea reproducible y escalable. Entre sus ventajas se incluyen:

- Automatización en la provisión de recursos cloud.
- Seguridad y consistencia en la configuración de servicios.
- Facilidad para escalar la infraestructura y realizar cambios.

III-B. Contenerización y Orquestación

Para la gestión y ejecución de la aplicación y la base de datos, se ha utilizado Docker[7] y Docker Compose[8], lo que permite encapsular las dependencias en contenedores y asegurar la portabilidad entre entornos.

- **Docker**: Para empaquetar y desplegar la aplicación.
- **Docker Compose**: Para la gestión de múltiples contenedores de forma automatizada.

III-C. Herramientas de Desarrollo y Administración

Se han utilizado diversas herramientas para el desarrollo y la gestión del entorno:

- **Python y Flask**: Framework para la aplicación web.
- **PostgreSQL**: Sistema de base de datos para almacenamiento de información.
- **PuTTY y SSH**: Acceso seguro a las instancias desplegadas en AWS.
- **DataGrip**: Herramienta para la gestión y consulta de bases de datos.

IV. Métodos

En el marco de este stack de formación, se presenta una estrategia que integra tanto la teoría como la práctica en la creación de infraestructuras seguras en la nube. La idea fue desarrollada a partir de experiencias personales en entornos de formación y se fundamenta en la adopción de herramientas y metodologías modernas, permitiendo no solo la implementación de un entorno robusto, sino también la consolidación de conocimientos clave en ciberseguridad y Cloud Computing.

IV-A. Requisitos de la Arquitectura

Antes de entrar en materia, es fundamental definir claramente los requisitos de la arquitectura para asegurar que el proyecto se alinee con los objetivos establecidos. Estos objetivos, previamente detallados en la seccion II, guían la creación de una arquitectura que cumpla con las necesidades del proyecto. Para lograr esto, es necesario fragmentar y analizar la petición de ViewNext, desglosando cada uno de los componentes y requisitos esenciales1.

A continuación, se detallan los componentes de la arquitectura:

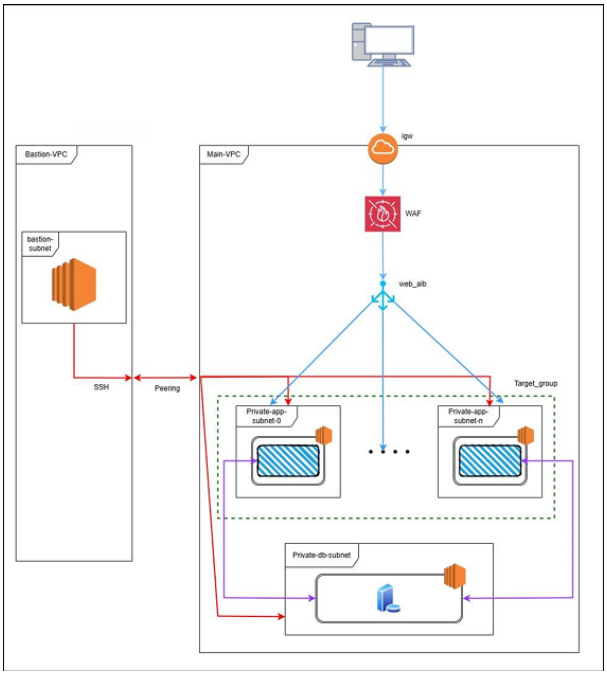


Figura 1. Diseño de la arquitectura a desplegar. Fuente: propia.

IV-A1. Bastion-VPC:

- **Bastion subnet**[9]: Contiene una instancia EC2 que actúa como bastion host, permitiendo acceso SSH seguro a las instancias en la Main-VPC. Esta VPC está conectada a la Main-VPC mediante peering.

IV-A2. Main-VPC:

- **Internet Gateway (IGW)**[10]: Permite la comunicación entre la VPC e internet.
- **Web Application Firewall**: Protege las aplicaciones web de ataques comunes de la capa 7, como inyecciones SQL o ataques de denegación de servicio (DDoS).
- **Application Load Balancer**: Distribuye el tráfico entrante entre las instancias del target group ubicadas en diferentes subredes privadas, asegurando la disponibilidad del servicio.
- **Private Web Subnet**[11]: Contiene instancias EC2 que ejecutan la aplicación web y reciben tráfico del ALB. Estas subredes están configuradas para no tener acceso directo a internet, garantizando un nivel adicional de seguridad.
- **Private Database Subnet**: Almacena la base de datos, accesible solo desde las subredes privadas donde residen las instancias EC2 que ejecutan la aplicación web. Esta configuración previene accesos no autorizados desde el exterior.
- **Target Group**: Grupo de instancias EC2 detrás del ALB que manejan el tráfico de la aplicación, asegurando una distribución equitativa de la carga de trabajo.

La implementación y desarrollo de esta arquitectura proporcionan una infraestructura que cumple rigurosamente con los requisitos definidos inicialmente. Una vez establecida la plantilla de diseño, se procede a ejecutar el plan acordado de manera sistemática y organizada.

IV-B. Diseño de la Arquitectura de la Web

El primer paso en la construcción de la infraestructura es el desarrollo de la aplicación web utilizando Flask, la cual se alojará en instancias EC2 en AWS. Una vez completada esta fase, se procederá con la configuración del balanceo de carga y la integración con la base de datos.

Se establecieron las siguientes funcionalidades clave para garantizar la operatividad de la aplicación:

- **Monitoreo del tráfico**: Dado que se utilizará un balanceador de carga, la aplicación deberá mostrar la ID del contenedor que está siendo accedido en cada momento, facilitando así la supervisión y resolución de problemas.
- **Conexión a la base de datos**: La aplicación debe conectarse a una base de datos para almacenar y gestionar información de manera eficiente.

Para asegurar el correcto funcionamiento de la aplicación, se han implementado las siguientes operaciones básicas:

- **Creación de elementos**: La aplicación cuenta con una interfaz que permite agregar nuevos registros a la base de datos mediante un formulario.
- **Visualización de datos**: Se proporciona una vista donde se listan todos los registros almacenados en la base de datos, actualizándose dinámicamente con cada modificación.
- **Eliminación de registros**: Se habilita una opción para borrar registros específicos de la base de datos a través de un icono de acción en la interfaz.

Estas funcionalidades básicas constituyen el núcleo de la aplicación web. Dado que el foco del proyecto está en la securización y despliegue en la nube, se ha optado por un diseño minimalista, priorizando la implementación de la infraestructura.

IV-C. Contenerización y Almacenamiento

IV-C1. Contenerización de la Aplicación Flask: Para garantizar la portabilidad y reproducibilidad de la aplicación web, se utiliza Docker para contenerizar la aplicación Flask. Este proceso implica la creación de un archivo Dockerfile que define la configuración necesaria para construir una imagen de contenedor que pueda ejecutarse de manera consistente en cualquier entorno.

El archivo Dockerfile incluye los siguientes pasos fundamentales:

- **Especificación de la imagen base**: Se elige una imagen base ligera de Python (Alpine Linux) para garantizar seguridad y eficiencia.
- **Creación de un usuario no root**: Se implementa una práctica recomendada de seguridad para ejecutar la aplicación con permisos restringidos.
- **Instalación de dependencias**: Se incluyen todas las librerías necesarias listadas en el archivo requirements.txt.
- **Definición del punto de entrada**: Se especifica el comando que ejecutará la aplicación Flask dentro del contenedor.

IV-C2. Configuración del Docker Compose: Para simplificar la gestión de contenedores, se utiliza Docker Compose, que permite definir y ejecutar múltiples contenedores de manera coordinada. La configuración incluye:

- **Definición del servicio web**: Se especifica la imagen de la aplicación Flask y se mapean los puertos correspondientes.
- **Aplicación de restricciones de seguridad**: Se implementan configuraciones como la eliminación de privilegios adicionales y la restricción del sistema de archivos a solo lectura.
- **Inicio automatizado**: Se establece un comando que permite iniciar todos los servicios con un solo comando.

IV-C3. Contenerización de la Base de Datos: La base de datos PostgreSQL se ejecuta en un contenedor separado utilizando una imagen preconfigurada. La configuración incluye:

- **Definición del servicio de base de datos**: Se establece la imagen de PostgreSQL y las variables de entorno necesarias.
- **Persistencia de datos**: Se configura un volumen para garantizar que los datos de la base de datos no se pierdan entre reinicios del contenedor.
- **Inicialización de datos**: Se monta un script SQL de inicialización para cargar la estructura de la base de datos automáticamente al iniciar el contenedor.

Este enfoque basado en contenedores garantiza que la aplicación y su base de datos puedan desplegarse de manera eficiente y segura en cualquier entorno, facilitando la escalabilidad y la administración de los servicios en la nube.

IV-D. Implementación en la Nube

Esta fase del proyecto se centra en la implementación de la arquitectura previamente diseñada en un entorno de computación en la nube, en este caso AWS. Antes de iniciar el despliegue, es fundamental configurar adecuadamente el entorno de trabajo para garantizar una implementación segura y eficiente.

IV-E. Configuración de Terraform

Antes de desplegar la infraestructura, se debe configurar Terraform para establecer una conexión segura con AWS, permitiendo la gestión de recursos de manera automatizada. Una correcta configuración de Terraform garantiza que la creación, actualización y eliminación de los recursos se realicen de forma estructurada y reproducible.

Terraform requiere la configuración de un proveedor, en este caso AWS, que define los parámetros de conexión. La especificación del proveedor se realiza de la siguiente manera:

- **Especificación del proveedor**: Se define AWS como el proveedor de infraestructura en la nube, indicando la versión del módulo necesario.
- **Configuración del proveedor**: Se establecen las credenciales de acceso a AWS, la región de despliegue y otros parámetros esenciales para la conexión.
- **Gestión de variables**: Se utilizan variables para almacenar información sensible, como claves de acceso y la región de despliegue, evitando que estos valores queden expuestos en el código fuente.

La gestión de variables se realiza a través del archivo `variables.tf`, que contiene las siguientes definiciones:

- **Clave de acceso y clave secreta**: Permiten la autenticación en AWS y deben almacenarse de manera segura.

- **Región de despliegue**: Define la ubicación de los recursos en AWS, asegurando disponibilidad y cumplimiento normativo.
- **AMI de Ubuntu 24**: Se especifica la imagen base que se utilizará para lanzar instancias EC2 en la nube.

Con esta configuración, Terraform está listo para iniciar el proceso de aprovisionamiento de la infraestructura en AWS, asegurando un despliegue controlado y escalable.

IV-F. Creación de AMIs Personalizadas

Para garantizar un despliegue optimizado, es necesario generar Amazon Machine Images (AMIs) personalizadas que contengan todos los componentes requeridos para ejecutar la aplicación web Flask y la base de datos. En este proceso, se ha utilizado Vagrant para definir y configurar las instancias antes de crear la AMI.

IV-F1. AMI para la Aplicación Web: Para la creación de la AMI[12] de la aplicación web, se siguieron los siguientes pasos:

- **Configuración del entorno**: Se utilizó Vagrant para definir una máquina virtual con Ubuntu 24, asegurando un entorno controlado y reproducible.
- **Instalación de dependencias**: Se instalaron Python3, Docker y Docker Compose para facilitar la ejecución de la aplicación Flask dentro de contenedores.
- **Contenerización de la aplicación**: Se ejecutó Docker Compose para desplegar y verificar el correcto funcionamiento de la aplicación antes de crear la AMI.
- **Configuración de red**: Se estableció la IP estática para la instancia, asegurando la conectividad con la base de datos en la infraestructura final.
- **Creación de la AMI**: Una vez verificado el funcionamiento del entorno, se procedió a la generación de la AMI a partir de la instancia configurada.

La AMI de la base de datos PostgreSQL se creó siguiendo un proceso similar, incluyendo la configuración de volúmenes persistentes y la ejecución de scripts de inicialización para garantizar que la base de datos estuviera lista para su uso inmediato.

Con la creación de estas AMIs, se optimiza el despliegue en AWS, reduciendo el tiempo de aprovisionamiento y asegurando que todas las instancias se inicien con la configuración predefinida y lista para operar.

IV-G. Creación de VPCs

El siguiente paso es la definición de la infraestructura de red mediante la configuración de la VPC principal (**main-VPC**), que albergará todos los recursos del proyecto.

- **Definición de la VPC**: Se establece un bloque CIDR de `10.0.0.0/16`, garantizando espacio suficiente para la segmentación de redes.
- **Subred de Base de Datos**: Se crea una subred privada en `10.0.1.0/24`, con una IP estática para el host de la BD.
- **Subredes de Aplicación**: Se configuran varias subredes privadas en distintas zonas de disponibilidad, permitiendo el balanceo de carga.
- **Gateway de Internet**: Se implementa un Internet Gateway (IGW) para la comunicación de la infraestructura con el exterior.
- **Tablas de enrutamiento**: Se establecen reglas de enrutamiento para permitir la conectividad entre los diferentes componentes de la VPC.

Para optimizar la escalabilidad, se emplea el atributo `count` en Terraform, lo que permite definir dinámicamente múltiples subredes en diferentes zonas de disponibilidad. Esta estrategia es crucial para distribuir las instancias de aplicación de manera equilibrada.

Finalmente, se asocian las subredes a una tabla de enrutamiento y se define una ruta predeterminada hacia el IGW, asegurando conectividad externa cuando sea necesario. Esta configuración completa la base de la infraestructura en la nube, garantizando seguridad y eficiencia en la implementación.

IV-H. Creación de Instancias EC2

Tras la configuración de la VPC, se procede con la creación de instancias para la base de datos y la aplicación web, asegurando su integración dentro de la infraestructura.

IV-H1. Instancia de la Base de Datos: Para desplegar la base de datos, se define un grupo de seguridad que restringe el acceso únicamente a las subredes privadas de la aplicación. La instancia EC2 correspondiente utiliza la AMI personalizada de la base de datos, se configura en una subred privada y se le asigna una IP estática para garantizar la conectividad.

IV-H2. Instancias de la Aplicación Web: Las instancias EC2 que alojarán la aplicación web siguen una configuración similar, con la diferencia de que pueden escalar en número según las necesidades del proyecto. Se implementan en subredes privadas y su tráfico es gestionado a través del balanceador de carga. Cada instancia está protegida por un grupo de seguridad que permite el tráfico HTTP y de Flask exclusivamente dentro de la VPC.

Esta configuración garantiza un entorno seguro y eficiente para el despliegue de la aplicación, optimizando la administración y escalabilidad de los recursos en la nube.

IV-I. Balanceador de Carga

Para garantizar la distribución eficiente del tráfico, se implementa un Application Load Balancer. Antes de su configuración, se define un **target group** que agrupa las instancias de la aplicación y gestiona el tráfico entrante.

- **Definición del Target Group**: Se configura para enrutar el tráfico HTTP en el puerto 5000 a las instancias de la aplicación.
- **Comprobaciones de salud**: Se establecen verificaciones periódicas para asegurar la disponibilidad de las instancias.
- **Asociación con instancias**: Se vinculan las instancias de la aplicación con el target group.

A continuación, se configura el ALB:

- **Definición del ALB**: Se implementa en múltiples zonas de disponibilidad para mejorar la resiliencia.
- **Reglas de tráfico**: Se establecen reglas de escucha en el puerto 80 que redirigen el tráfico al target group.
- **Grupo de seguridad**: Se permite tráfico entrante en el puerto 80 desde cualquier origen.

Esta configuración asegura una distribución uniforme del tráfico y permite escalar la aplicación de forma eficiente.

IV-I1. Configuración del AWS WAF: Se implementan reglas de seguridad para restringir accesos no totorizados[13]:

- **Restricción de tráfico**: Se bloquea el acceso desde la VPC de bastion y se implementa un CAPTCHA para accesos externos.
- **Monitoreo de tráfico**: Se habilitan métricas de CloudWatch para analizar patrones de acceso.
- **Integración con ALB**: Se asocia el ACL del WAF con el ALB para filtrar tráfico malicioso.

Esta configuración asegura una distribución uniforme del tráfico y permite escalar la aplicación de forma eficiente, al mismo tiempo que protege la infraestructura de ataques y accesos no deseados.

V. Resultados

En este capítulo se presentan los resultados obtenidos tras la finalización del desarrollo de la arquitectura. A través de ilustraciones, salidas de comandos y pruebas en tiempo real, se demuestra cómo cada componente implementado se integra en un entorno práctico de formación, permitiendo validar tanto la robustez como la funcionalidad del sistema.

La implementación abarcó desde la creación de los recursos con Terraform hasta la verificación de la conectividad y seguridad de la infraestructura.

V-A. Creación de los Recursos Definidos con Terraform

Para verificar la correcta definición de los recursos se ejecutó el comando `terraform apply`. Antes de proceder, se establecieron variables clave, como el número de instancias a desplegar y la IP personal.

La ejecución de `terraform apply` generó un plan preliminar similar a:

```
Plan: 43 to add, 0 to change, 0 to destroy.
```

y, tras su aprobación, se obtuvo la confirmación:

```
Apply complete! Resources: 43 added, 0 changed, 0 destroyed.
```

Esta salida confirma que todos los recursos (como instancias EC2 y subredes numeradas) se crearon correctamente, proporcionando un entorno coherente para la formación práctica.

V-B. Acceso a la Aplicación Web y Conexión con la Base de Datos

Una vez desplegados los recursos, se verificó el acceso a la aplicación web a través de la URL configurada (`tfgawsbadre.duckdns.org`). Al ingresar, el sistema mostró inicialmente una pantalla de verificación CAPTCHA, demostrando la correcta implementación de la regla ACL que solicita la validación de humanidad (véase la Figura 2).

Posteriormente, al acceder a la página principal, se observa que la aplicación desplegada en Flask mostraba el identificador del contenedor activo y ofrecía funcionalidades para añadir, visualizar y eliminar elementos de la base de datos. Se realizó una prueba de inserción de un registro (por ejemplo, una nueva fruta) y se comprobó que el cambio se reflejara en la lista de elementos, confirmando la integración con PostgreSQL (ver Figura 8).

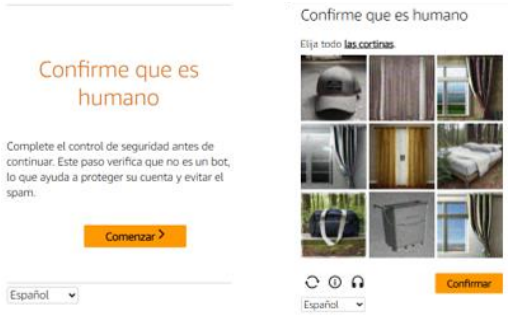


Figura 2. Pantalla de verificación CAPTCHA en la aplicación web. Fuente: propia.

V-C. Balanceo de Carga

El Application Load Balancer demostró su capacidad para distribuir el tráfico entre las instancias de la aplicación. Al refrescar la página, se observaron distintos identificadores de contenedores en cada solicitud, lo que evidencia que el tráfico se está equilibrando de forma efectiva (ver Figura 3). Aunque la cantidad de peticiones fue limitada en este entorno de formación, la variabilidad de identificadores respalda el funcionamiento del balanceo.

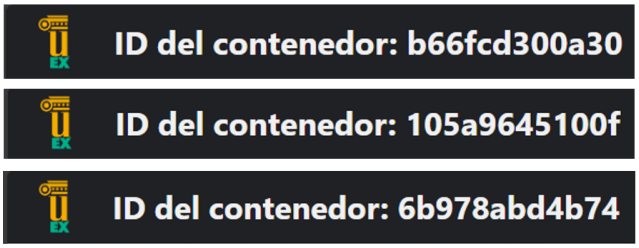


Figura 3. Identificadores de contenedores en la aplicación web, evidenciando el balanceo de carga. Fuente: propia.

V-D. Administración Segura

La seguridad y administración remota se validaron mediante el acceso SSH a través del host bastión, utilizando PuTTY. Al conectarse con la IP pública del bastión y la clave privada, se pudo acceder a la instancia, confirmándose que solo la IP personal autorizada permitía el acceso. Se realizaron pruebas con VPN, y los intentos de conexión desde IPs distintas fueron rechazados (véase las Figuras 5 y 4).

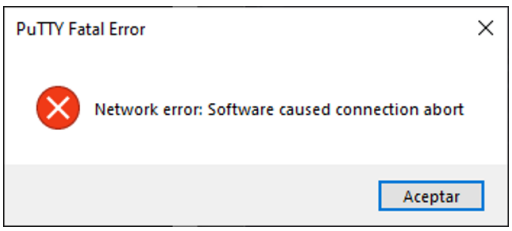


Figura 4. Mensaje de error en PuTTY al intentar conectarse desde una IP no autorizada. Fuente: propia.

Además, mediante el comando `docker ps` se corroboró que el contenedor de la aplicación (basado en la imagen `flask-app-image`) se encontraba activo y operativo (ver Figura 6).

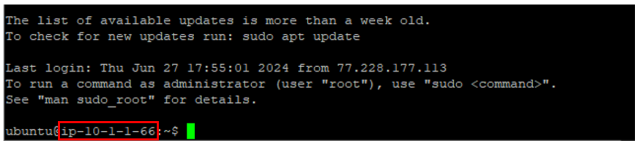


Figura 5. Conexión SSH exitosa a la instancia de bastión utilizando PuTTY. Fuente: propia.

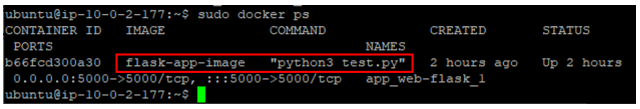


Figura 6. Salida del comando `docker ps` mostrando la actividad del contenedor de la aplicación. Fuente: propia.

V-E. Validación de Reglas ACL y Control de Acceso

Se comprobó el funcionamiento de las reglas ACL implementadas en AWS WAF. Con la herramienta `curl` se realizó una solicitud desde una instancia no autorizada, obteniéndose un error `403 Forbidden` (ver Figura 7). Este comportamiento confirma que la política de bloqueo configurada impide el acceso desde orígenes no permitidos.

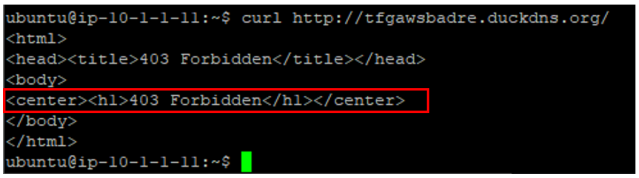


Figura 7. Salida del comando `curl` mostrando el error 403 Forbidden. Fuente: propia.

V-F. Destrucción de la Infraestructura

Finalmente, para evitar costos adicionales y concluir el proceso formativo, se ejecutó el comando `terraform destroy`. Este comando presentó un plan de destrucción, y tras su confirmación, se eliminaron todos los recursos creados, como se evidencia en el mensaje:

```
Destroy complete! Resources: 43 destroyed.
```

Esta acción demuestra la capacidad de gestionar el ciclo de vida completo de la infraestructura.

VI. Conclusiones

La implementación de esta infraestructura segura en AWS utilizando Terraform ha evidenciado las ventajas del uso de la Infraestructura como Código en términos de automatización, seguridad y eficiencia operativa. La posibilidad de definir y desplegar entornos completos de manera reproducible ha permitido estandarizar procesos, reducir errores y mejorar el control de acceso a los recursos.

VI-A. Principales aportaciones

- **Análisis de requerimientos de seguridad**: Se realizó una evaluación minuciosa de las necesidades de seguridad en AWS, identificando las mejores prácticas y estrategias recomendadas para mitigar riesgos y garantizar la protección de los datos y servicios desplegados.
- **Diseño de infraestructura segura**: A través de la implementación de VPCs, subredes privadas y balanceadores de carga, se logró una arquitectura modular que refuerza la seguridad y permite una gestión eficiente de los recursos.
- **Automatización del despliegue**: La utilización de Terraform facilitó el aprovisionamiento automatizado de la infraestructura, garantizando un despliegue replicable, escalable y alineado con los estándares de seguridad.
- **Integración de servicios y redes**: Se estableció una comunicación fluida entre los distintos componentes de la arquitectura, permitiendo una interoperabilidad eficiente entre los servicios de aplicación y base de datos.
- **Implementación de mecanismos de protección**: Se integró AWS WAF como un mecanismo de protección para el balanceador de carga, fortaleciendo la defensa contra amenazas y accesos no autorizados.

VI-B. Posibles Ampliaciones y Líneas Futuras de Trabajo

Con respecto a las líneas futuras, se proponen varias mejoras específicas que permitirán evolucionar el sistema presentado:

- **Expansión multicloud**: Se podría desarrollar módulos adicionales específicos para otros proveedores cloud (Azure y Google Cloud Platform) en Terraform, realizando pruebas y evaluaciones comparativas para validar su interoperabilidad y eficacia en distintos entornos.
- **Fortalecimiento de la seguridad**: Se podría realizar mediante la integración directa de servicios avanzados como AWS GuardDuty (configurando alertas automatizadas frente a actividades sospechosas), AWS Macie (estableciendo reglas automatizadas de monitoreo para datos sensibles) y AWS Cognito (implementando flujos de autenticación y gestión avanzada de identidades).
- **Automatización avanzada de gestión de secretos**: Se podrían implementar herramientas especializadas como AWS Secrets Manager o HashiCorp Vault mediante el desarrollo de scripts automatizados específicos, que faciliten la gestión segura y centralizada de credenciales y secretos operativos, con revisiones periódicas de seguridad.
- **Optimización continua del rendimiento**: Podría realizarse mediante la implementación de herramientas de monitoreo avanzadas como AWS CloudWatch, configurando alertas automatizadas para detectar y solucionar proactivamente cuellos de botella, y ajustando recursos de forma dinámica con políticas de escalado automático basadas en métricas en tiempo real.
- **Incorporación con tecnología emergentes**: Se podría realizar desarrollando modelos específicos de detección de anomalías usando servicios como AWS SageMaker, permitiendo identificar rápidamente actividades inusuales y responder de manera automatizada mediante integraciones con herramientas de gestión de incidentes.
- **Evaluación continua mediante pruebas avanzadas**: Se podrían implementar mediante la realización periódica de pruebas de penetración avanzadas utilizando metodologías específicas (por ejemplo, OWASP, PTES), colaborando con equipos especializados (Red Team) para simular ataques realistas y evaluar la resistencia de la

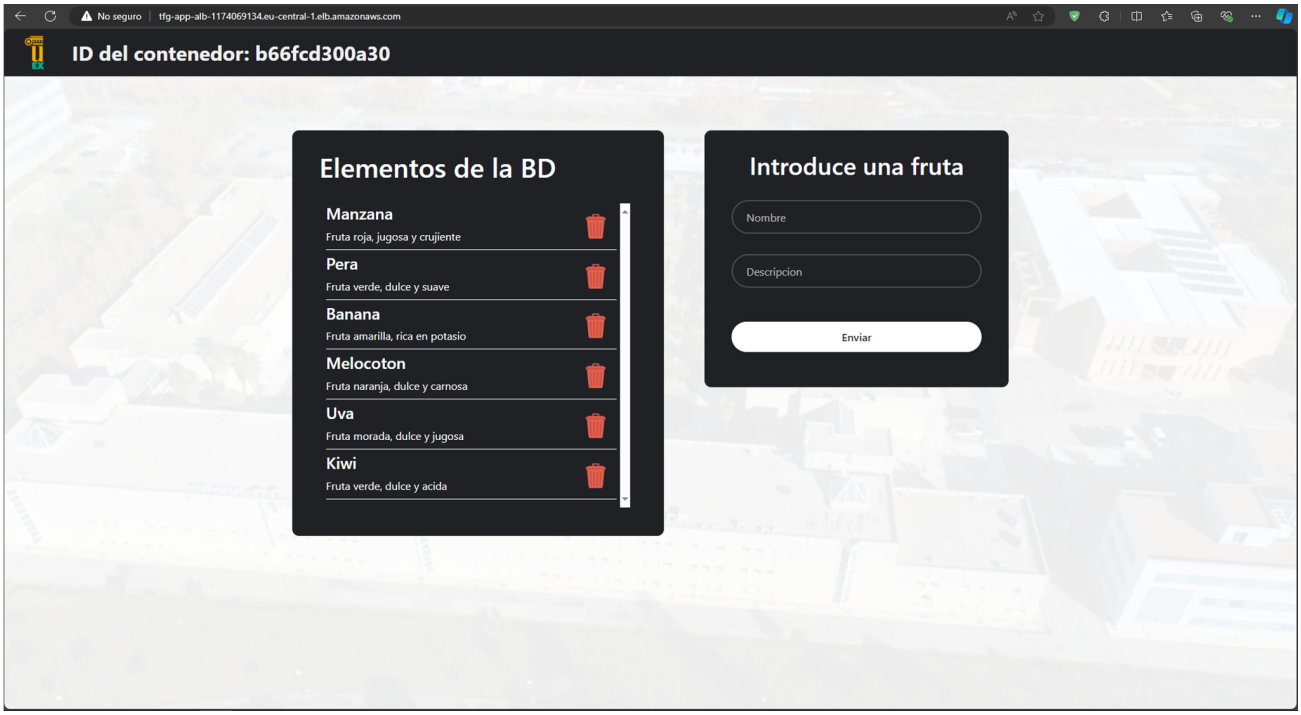


Figura 8. Vista de la aplicación web mostrando la lista de elementos en la base de datos. Fuente: propia.

infraestructura, proporcionando informes detallados para mejorar continuamente las defensas.

Estas mejoras permitirán ampliar la robustez y eficiencia de la infraestructura, manteniendo su utilidad educativa y asegurando una adaptación constante a las nuevas amenazas y tecnologías emergentes en el campo de la ciberseguridad.

VII. Agradecimientos

Esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation) - Instituto Nacional de Ciberseguridad (INCIBE) en el proyecto C109/23 ”Proyecto estratégico UEx (Escuela Politécnica de Cáceres)-INCIBE”.

Referencias

[1] “Seguridad de la infraestructura de Amazon EC2 - Amazon Elastic Compute Cloud — docs.aws.amazon.com,” https://docs.aws.amazon.com/es_es/AWSEC2/latest/UserGuide/infrastructure-security.html, [Accessed 07-03-2025].

[2] “&xBF;Qu&xE9; es Amazon VPC? - Amazon Virtual Private Cloud — docs.aws.amazon.com,” https://docs.aws.amazon.com/es_es/vpc/latest/userguide/what-is-amazon-vpc.html, [Accessed 28-03-2025].

[3] “Control traffic to your AWS resources using security groups - Amazon Virtual Private Cloud — docs.aws.amazon.com,” https://docs.aws.amazon.com/vpc/latest/userguide/vpc-security-groups.html, [Accessed 28-03-2025].

[4] S. Ninawe, “Managing Application Load Balancer (ALB) with Terraform — spacelift.io,” https://spacelift.io/blog/terraform-alb#prerequisites--the-target-architecture, 2024, [Accessed 07-03-2025].

[5] “Target groups for your Application Load Balancers - Elastic Load Balancing — docs.aws.amazon.com,” https://docs.aws.amazon.com/elasticloadbalancing/latest/application/load-balancer-target-groups.html, [Accessed 28-03-2025].

[6] C. Cilleruelo, “¿Qué es AWS WAF? — KeepCoding Bootcamps — keepcoding.io,” https://keepcoding.io/blog/que-es-aws-waf/, 2024, [Accessed 07-03-2025].

[7] “¿Qué es Docker y cómo funciona? Ventajas de los contenedores Docker — redhat.com,” https://www.redhat.com/es/topics/containers/what-is-docker, 2023, [Accessed 07-03-2025].

[8] A. B, “What is Docker Compose? How to Use it with an Example — freecodecamp.org,” https://www.freecodecamp.org/news/what-is-docker-compose-how-to-use-it/, 2023, [Accessed 07-03-2025].

[9] Dhfernando, “Bastion Host in AWS VPC — dev.to,” https://dev.to/aws-builders/bastion-host-in-aws-vpc-2i63, 2023, [Accessed 07-03-2025].

[10] “Enable internet access for a VPC using an internet gateway - Amazon Virtual Private Cloud — docs.aws.amazon.com,” https://docs.aws.amazon.com/vpc/latest/userguide/VPC_Internet_Gateway.html, [Accessed 28-03-2025].

[11] “Subnets for your VPC - Amazon Virtual Private Cloud — docs.aws.amazon.com,” https://docs.aws.amazon.com/vpc/latest/userguide/configure-subnets.html, [Accessed 28-03-2025].

[12] “Im&xE1;genes de m&xE1;quina de Amazon (AMI) en Amazon EC2 - Amazon Elastic Compute Cloud — docs.aws.amazon.com,” https://docs.aws.amazon.com/es_es/AWSEC2/latest/UserGuide/AMIs.html, [Accessed 07-03-2025].

[13] K. Team, “Configure AWS WAF With Terraform — KubeRocketCI — docs.kuberocketci.io,” https://docs.kuberocketci.io/docs/operator-guide/infrastructure-providers/aws/waf-tf-configuration, [Accessed 07-03-2025].

Hybrid Tabletop Exercise (TTX) based on a Mathematical Simulation-based Model for the Maritime Sector

Diego Cabuya-Padilla[1], Daniel Díaz-López[2,3], Carlos Castaneda-Marroquín[1]
[1] "Cyberspace, Technology and Innovation" research group, Escuela Superior de Guerra "General Rafael Reyes Prieto", Cra. 11 # 102-50, Bogotá, Colombia
[2]Department of Information and Communications Engineering, University of Murcia, 30100, Murcia, Spain
danielorlando.diaz@um.es
[3]School of Engineering, Science and Technology, Universidad del Rosario, Bogotá, Colombia
danielo.diaz@urosario.edu.co

***Abstract*—As cyber threats grow in complexity and scale, many security incidents remain poorly managed due to the lack of proper training among C-level executives. Thus, there is a need for targeted cybersecurity education to enhance executive decision-making and crisis response. Traditional training methods, such as cyber wargames and Tabletop Exercises (TTX), aim to develop abilities to face critical incidents, however, they often lack the interactive and dynamic elements required to prepare individuals for real-world cyber incidents. This paper presents a novel approach to cybersecurity and cyberdefense education through the design of a specialized hybrid TTX for the maritime domain, which uses a framework to model mathematically how a cyberattack spreads along multiple nodes and impacts infrastructure. Our proposal was validated through exercises in Argentina and the United States, demonstrating a positive impact in developing the comprehension and projection levels of Cyber Situational Awareness (CSA), and reinforcing governance. Documentation about the Hybrid TTX, scenario, datasets and implementation of the SERDUX-MARCIM model, is available at the project repository at https://github.com/diegocabuya/SERDUX-MARCIM**

***Index Terms*—Cybersecurity, Cyberdefense, Maritime, Serious Games, Tabletop Exercises (TTX), Cyber Situational Awareness (CSA), Education in cybersecurity**

Contribution type: *Training and educational innovation*

I. Introduction

Tabletop Exercises (TTX) are structured discussion-based activities designed to train teams in responding to cybersecurity incidents and crisis situations [1]. Unlike hands-on technical training, TTX focuses on decision-making, communication, and coordination under simulated emergency conditions. These exercises provide a controlled environment where participants assume organizational roles and collaboratively address evolving cyber threats through scenario-driven discussions [2]. TTX's are widely used in cybersecurity training and risk management, as they help organizations evaluate their response capabilities and identify gaps in policies and procedures. Their adaptability and cost-effectiveness make them a valuable tool for improving resilience against cyberattacks while fostering a culture of proactive security awareness [3].

In addition, Cyber Situational Awareness (CSA) is a fundamental element, as it enables organizations to understand and anticipate cyber threats by processing and analyzing relevant information from various sources [4]. CSA is based on the broader concept of situational awareness, which consists of 3 levels: perception, comprehension, and projection [5]. The perception level involves identifying key cyber elements, such as anomalies in network traffic or security alerts. The comprehension level integrates and analyzes this information to determine its significance, identifying potential threats or vulnerabilities. Finally, the projection level allows decision makers to anticipate the evolution of cyber incidents, assessing their potential impact, and determining proactive response strategies. These levels collectively improve an organization's ability to detect, assess, and mitigate cyber risks, forming a crucial foundation for proactive cybersecurity.

TTX establishes a structured and controlled scenario that allows participants to analyze decision-making processes and response strategies to cyber incidents [6]. However, traditional TTX methodologies face significant challenges, such as limited real-time data integration, lack of dynamic adversarial behavior modeling, and difficulties in quantitatively assessing decision-making [7]. Thus, to address the challenges identified in traditional TTX methodologies, this paper proposes an evolution towards a hybrid TTX approach that integrates serious games and simulation-based techniques. The main contributions of this paper are summarized as follows:

- Design of a novel Hybrid TTX Methodology that integrates the traditional TTX approach with a mathematical simulation-based model (SERDUX-MARCIM).
- Application of the proposed hybrid TTX methodology in the maritime sector, strengthening CSA levels among strategic, including military personnel, maritime authorities, and policymakers.
- Evaluation of our proposal in 2 real-World Scenarios, one conducted in Argentina, and the other in United States.

This paper is structured as follows: Section III introduces

the SERDUX-MARCIM model. Section IV describes our proposal. Section V and Section VI details the execution of two TTX exercises, highlighting participant insights on CSA. Section VII presents the conclusions and outlines future work.

II. State of the art

A bibliometric review reveals a lack of specialized tools to support the strategic adoption of response protocols in maritime cybersecurity and cyberdefense. Nonetheless, some studies offer valuable insights for the proposal.

Damodaran and Wagner [8] highlight simulation's role in cyberdefense through game-theoretic models and network recovery strategies, stressing its value for cyber situational awareness (CSA). Similarly, Bodeau et al. [9] and Katsantonis [10] present wargaming frameworks linking cybersecurity training to risk management. Valente and Reith [11] advocate serious gaming over traditional training in military settings, emphasizing the inclusion of IT and OT systems to improve awareness in critical infrastructure. Onduto [12] supports gamification's impact on behavior but warns of its experimental maturity and limited assessment of long-term learning.

In summary, simulation-based methods have enhanced cybersecurity education, but critical gaps persist in the maritime sector. Current tools often fail to capture the complexity of interconnected maritime systems, signaling the need for novel, simulation-driven methodologies to reinforce strategic preparedness and structured cyber response.

III. Background

As part of the MARCIM research project [13], SERDUX-MARCIM [14] was developed as a mathematical and simulation model to support experimenters in formulating and testing working hypotheses, test courses of action, and identifying potential maritime cybersecurity and cyberdefense scenarios.

SERDUX-MARCIM combines three techniques: compartmental epidemiological models, system dynamics, and agent-based modeling. This combination allows for a more comprehensive understanding of how cyberattacks propagate within a maritime ecosystem, simulating the attack as a dynamic process. By adapting epidemiological models such as SIR (Susceptible, Infected, Recovered) and SEIR (Susceptible, Exposed, Infected, Recovered) to the context of cyberattack propagation, this model offers a way to assess risk and predict the behavior of such attacks as they unfold, making it relevant for both operational validation and strategic awareness.

Finally, SERDUX-MARCIM enhances the three strategic levels of CSA, i.e. perception, comprehension, and projection, by offering a framework that guides decision-making during cyberattacks on maritime infrastructure, improving response coordination and resilience in maritime cybersecurity.

IV. Hybrid Tabletop Exercise Methodology for Maritime Sector

A Hybrid TTX methodology is proposed in this section integrating the traditional TTX approach [15] with the SERDUX-MARCIM model [14], introduced in Section III. By incorporating serious gaming techniques, this approach allows participants to interact with the computational model, gaining deeper insights into its implications within a controlled yet realistic decision-making environment.

In this context, the hybrid TTX is designed to strengthen maritime cybersecurity and cyberdefense by providing a high-level perspective. A key emphasis is placed on CSA as a fundamental component of informed decision-making across strategic, tactical, and operational levels.

A. Scenario

Participants worked with the 2017 NotPetya cyberattack against Maersk [16], [17], [18] as a reference case to implement a case-based learning strategy aimed at developing CSA levels. This scenario was selected due to its relevance as one of the most disruptive cyber incidents in the maritime sector, exposing critical vulnerabilities in maritime infrastructures [14]. By using this cyberattack, the exercises allowed participants to reflect on the complexities of cyber threats in the maritime domain and the importance of comprehensive risk assessments, as well as the design of effective strategies.

Within the SERDUX-MARCIM framework, an Event i represents a key milestone in the progression of a cyber crisis scenario and it belongs to any of the 3 standard crisis phases [18]: pre-crisis, crisis, and post-crisis. Thus, for the application of the SERDUX-MARCIM framework in a cyberdefense scenario, we must define the key variables indicated in Table I. To apply SERDUX-MARCIM in the Maersk case, we consider the execution of 3 simulations, i.e. $s = 3$, 4 key events along the scenario, i.e. $i = 4$, and 3 variables to monitor, i.e. $pm = 3$. Specifically, the variables to monitor in this scenario are i) network situation, ii) level of service availability, and iii) overall cyber risk posture. This setup is illustrated in Figure 1. This structured approach allows participants to systematically evaluate how cyber threats evolve and how decision-making dynamics shift throughout a crisis.

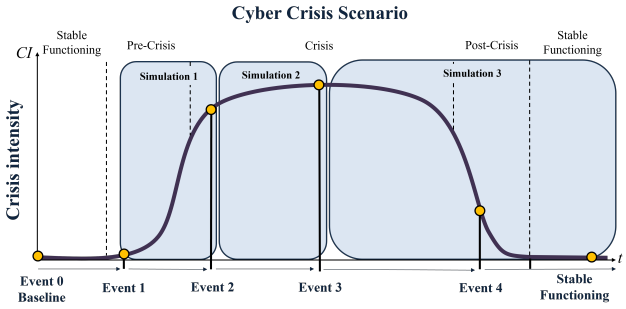


Figure 1: Simulations and events defined for the application of SERDUX-MARCIM framework in the Maersk case [14]

B. Methodology

The proposed Hybrid TTX methodology consists of seven steps, incorporating a structured decision-making process supported by the SERDUX-MARCIM model. This hybrid ap-

proach enhances CSA at a strategic level by engaging participants in simulated cyber crisis scenarios. While the methodology contributes to the development of all CSA levels, it primarily strengthens the comprehension and projection skills of participants, allowing them to interpret complex situations and anticipate potential outcomes. The proposed methodology, along with its phases and steps, is depicted in Figure 2.

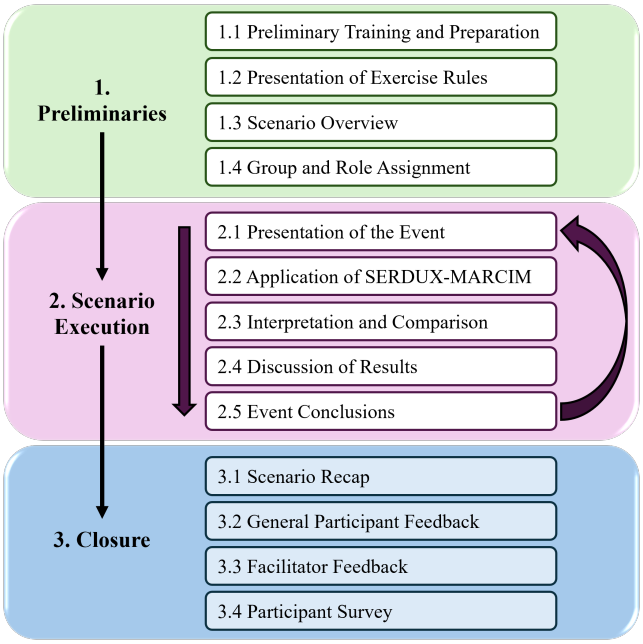


Figure 2: Methodology proposed for the hybrid TTX

The preliminary phase is designed to prepare participants and establish the conditions for the successful execution of the exercise. Its main purpose is to standardize participants' knowledge, define the rules and structure of the activity, and ensure that everyone clearly understands the scenario and their assigned roles. This phase lays the groundwork for meaningful discussions and informed decision-making during the exercise.

The scenario execution phase consists of the dynamic development of the cybersecurity scenario through sequential events. Each event follows a cyclical process, repeating as many times as the scenario establishes. Each cycle is composed of five steps describe below:

- **Presentation event** i: Event i is presented to participants, providing relevant details about it: organizational context, key actors involved, and any critical information necessary to understand the situation.
- **Application of the SERDUX-MARCIM**: The event i is integrated in the SERDUX-MARCIM model, which provides a structured cyber risk analysis. This allows participants to dynamically assess the evolution of the attack in terms of impact on maritime infrastructure.
- **Interpretation and comparison**: Participants analyze the outputs generated by the SERDUX-MARCIM model, focusing on understanding the current situation (comprehension) and anticipating potential future developments (projection) of the cyberattack.
- **Discussion of results**: Facilitators use carefully designed guiding questions to promote a structured discussion aimed at strengthening participants' CSA, with a particular focus on developing their Projection capabilities.
  - First, participants are invited to analyze the current situation of the event, reflecting on the implications for the maritime sector, identifying affected services, and assessing the role of internal and external actors involved.
  - Second, the discussion shifts toward the formulation of possible courses of action. Participants evaluate response alternatives based on the current scenario dynamics and the projected impacts of their decisions as estimated by the model outputs.
  - Finally, participants are guided to deliberate and agree on the most favorable course of action.
- **Event conclusions**: The group leader presents the most relevant insights and conclusions based on the guiding questions. These conclusions summarize key takeaways, lessons learned, and potential improvements in cybersecurity and cyberdefense strategies.

The closure phase is in charge of consolidating the key lessons learned, evaluating the exercise outcomes, and gathering participant feedback to strengthen future training activities. This phase aims to promote collective reflection on the decisions made, the strategies discussed, and the overall development of CSA throughout the exercise.

V. Experiments

This section describes the experiments conducted in Buenos Aires, Argentina, and Washington D.C, United States, where the hybrid TTX proposed in Section IV was applied to assess its effectiveness in fostering CSA at the strategic level.

A. Participants

The exercises involved 36 high-level international stakeholders with strategic responsibilities in maritime security. Participants' roles included military personnel, maritime authorities, policymakers, deputy ministers, ambassadors, diplomats, academics, and cybersecurity specialists. The selection criteria prioritized individuals with extensive experience in strategic decision-making and crisis management. Notably, given the strategic focus of the exercise, participants were not undergraduate or graduate students, but professionals actively engaged in roles tied to maritime power. Gathering such a distinguished and diverse group, spanning 16 countries, mainly from the Americas, is particularly significant; the sample size is substantial considering the highly specific profile required, which makes this level of participation exceptionally rare in conventional training environments.

B. Exercises Description

1) Exercise 1: It was conducted in Buenos Aires, Argentina, on February 23, 2024, during the *SALMA Dialogue:*

For a More Efficient Maritime Security in the Central and South Atlantic. The variables considered for the exercise are listed in Table I. The methodology involved presenting SERDUX-MARCIM model simulations applied to the Maersk cyberattack within a structured TTX format. This approach integrated plenary sessions for scenario analysis and small group discussions (3–4 participants), allowing in-depth examination of key cybersecurity challenges. Following the exercise, participants completed an individual survey, which will be described in Section V-C, to assess their experience and provide feedback on the model's applicability.

Variable	Symbol	Exercise 1	Exercise 2
Number of Simulation	s	3	3
Number of events to consider	i_t	5	5
Perspectives to monitor per event	pm	3	3
Number of Participants	np	10	26
Number of Observers	no	15	4
Group Size	gs	3-4	8-12

Table I: Variables defined per exercise

2) Exercise 2: It took place in Washington, D.C., United States, from May 16 to 17, 2024, as part of the *Advanced Course on Maritime Security Policies*, organized by the William J. Perry Center at the National Defense University, U.S. The variables considered for the exercise are related in Table I, and the exercise followed a similar methodology to Exercise 1.

C. Data Collection Method

According to the methodology outlined in Section IV, the final step of the *Closure* phase involves the application of a *Participant Survey*. Accordingly, two surveys were conducted, comprising 18 variables [X1–X18] as detailed in Table II. These variables were selected to ensure a comprehensive evaluation of the exercise, encompassing aspects related to the participants, the exercise execution, and the scenario design.

- **Survey 1:** TTX assessment, structured into three sections:
  1) Participants' previous experience with TTX or simulations, particularly in the maritime domain, as well as their training in cybersecurity or cyberdefense. [X1-X4]
  2) General evaluation of the exercise. [X5-X8]
  3) Scenario evaluation. [X9-X13]
- **Survey 2:** SERDUX-MARCIM model overall accuracy. [X14-X18]

Data collected through these surveys was analyzed to identify statistical associations and performance patterns, providing insights into participants' perceptions regarding the effectiveness of the exercise, the applicability of the SERDUX-MARCIM model in real-world scenarios, and the appropriation of CSA skills (Perception, Comprehension, Projection) at the strategic level. To ensure data protection, sensitive participant information, such as names and affiliations, was omitted.

D. Survey Results

1) Survey 1 - TTX Assessment:

a) Section 1 - Participants' Prior Experience and Cybersecurity or Cyberdefense Training: Findings (Figure 3) indicate that only 55.6% of the participants had previously participated in a TTX or a crisis simulation exercise ($X1$). Additionally, 13.9% had taken part in cybersecurity and cyberdefense simulation exercises ($X2$), and 11.1% had participated in maritime cybersecurity and cyberdefense simulation exercises ($X3$). Moreover, 52.8% of the participants had received some formal or informal training in cybersecurity ($X4$).

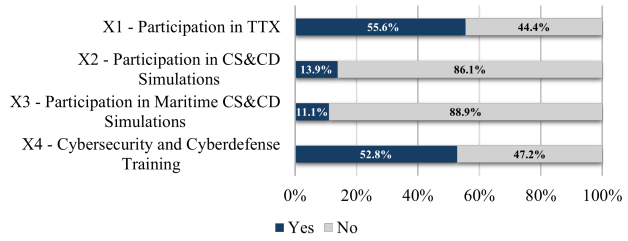


Figure 3: Survey 1 - Participants' Prior Experience and Training [0% - 100%]

The analysis highlights a significant lack of engagement in cybersecurity simulation exercises, particularly in the maritime sector. Furthermore, the results emphasize the significant need for enhanced training, especially at the strategic level, to strengthen maritime cybersecurity preparedness.

b) Section 2 - General Evaluation of the Exercise: Figure 4 depicted that the exercises received a highly favorable rating, with an average score of 4.56/5, considering aspects such as facilitator effectiveness ($X5$), time management ($X6$), resources ($X7$), and overall participant satisfaction (Y). Participants expressed a high level of satisfaction and strong interest in replicating such exercises in their respective countries to improve cyber awareness and crisis management capabilities. Additionally, the exercises facilitated the internalization of fundamental concepts related to the typical evolution of cyber crises and the strategic responses required at each phase.

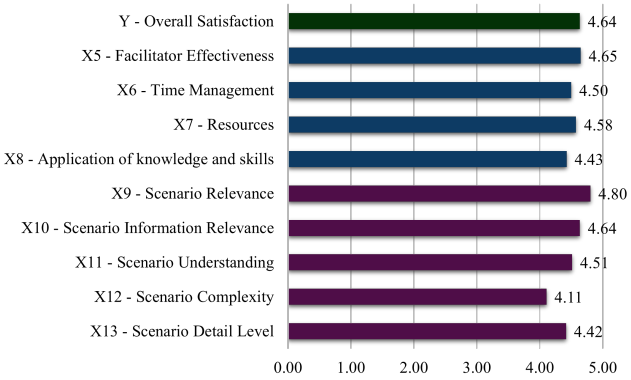


Figure 4: Survey 1 - General Evaluation - Scenario Evaluation [0 - 5]

Despite the positive outcomes, participants suggested additional preparatory materials to optimize engagement, extend-

ing the duration of the exercise, and introducing a full-day cybersecurity and cyberdefense briefing beforehand.

c) Section 3 - Scenario Evaluation: Figure 4 depicted that the scenario evaluation was also highly positive, with an average rating of 4.52/5. The assessed criteria included questions about different aspects of the proposed cyberdefense scenario: relevance ($X8$), information relevance ($X9$), understanding ($X10$), complexity ($X11$), level of detail ($X12$), and application of knowledge and skills ($X13$). Participants concluded that the scenario was highly relevant to the maritime sector, and the provided information was adequate.

However, due to the varying levels of prior cybersecurity knowledge among participants, some found the scenario somewhat complex, particularly regarding the technical aspects of the final events. As an area for improvement, participants recommended providing a clearer strategic interpretation of the tactical and operational situation during these events.

2) Survey 2 - SERDUX-MARCIM Model Overall Accuracy: The SERDUX-MARCIM model assessment (Figure 5) received highly positive feedback, with an average score of 4.67 out of 5 across key evaluation criteria: relevance ($X14$), accuracy ($X15$), the need for similar models in other organizations ($X16$), integration potential ($X17$), and expected applicability within the intended domain ($X18$).

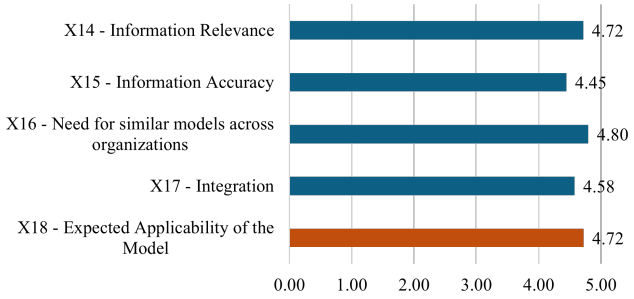


Figure 5: Survey 2 - SERDUX-MARCIM model overall accuracy [0 - 5]

These results underscore SERDUX-MARCIM's effectiveness as a tool for enhancing maritime cybersecurity awareness, particularly by improving the comprehension and projection dimensions of CSA. This improvement is supported from two perspectives. First, the high perceived value that participants assigned to the model-based outputs, especially regarding relevance ($X14$) and accuracy ($X15$), which aligns with the TTX objectives. Second, SERDUX-MARCIM's ability to provide participants with actionable information that enables the analysis of the current situation, the observation of variable trends over time, and the visualization of potential future behaviors. Specifically, the model facilitates the examination of three key elements: (i) the network situation, (ii) the level of service availability, and (iii) the overall cyber risk posture. This comprehensive perspective allowed participants to better understand the evolving dynamics of the simulated cyberattack and to anticipate its potential impacts on maritime operations.

Moreover, SERDUX-MARCIM demonstrates its potential as a decision-support tool in strategic responses to cyber threats, as reflected in the positive evaluations of the need for similar models across organizations ($X16$) and its integration capabilities ($X17$).

Finally, the results for the expected applicability of the model ($X18$) confirm that SERDUX-MARCIM achieves the required accuracy for its intended purpose and is well-suited for maritime cybersecurity and cyberdefense contexts. Participants particularly emphasized the model's practical utility and the effectiveness of the tools and methods employed.

Additionally, regarding the general observations ($X19$) gathered through an open question, participants acknowledged that integrating models like SERDUX-MARCIM into their organizations could present certain challenges. However, there was significant interest in adopting similar initiatives. These findings reinforce the validity of the proposed framework and highlight its potential to strengthen cybersecurity and cyberdefense capabilities within critical maritime infrastructure.

VI. Prospective Analysis

The objective of this section is to provide a comprehensive prospective analysis aimed at identifying areas for improvement in the design and execution of future cybersecurity and cyberdefense training exercises, particularly within the maritime sector. This analysis is divided into two main sections: First, *Multiple Linear Regression Model*, utilizes statistical modeling to predict the influence of key variables on participant outcomes. Finally, the *Prospective Scenario Analysis* explores potential future scenarios—pessimistic, trend-based, and optimistic—to guide the refinement of future exercises based on the analysis conducted. Each section contributes to a deeper understanding of how to optimize the learning and engagement of participants in future exercises.

A. Objective and Methodology

To gain insights for improving future iterations of the proposed hybrid TTX, a mathematical analysis was conducted based on Survey 1 (Section V-D1). This analysis aimed to assess the influence of attributes related to i) participants, ii) TTX execution, and iii) presented scenario, on the overall satisfaction with the TTX and in the ability to apply the knowledge acquired during the TTX.

Two analytical approaches were applied: i) multiple linear regression modeling, and ii) prospective scenario analysis. The dependent variable for these analyses was "Overall Satisfaction" (Y), while independent variables included participant experience ($X1 - X4$) and exercise perception ($X5 - X13$).

B. Multiple Linear Regression Model

The multiple linear regression model developed for the collected data is expressed in Equation 1.

$$\begin{aligned} Y = &-0.0252X1 + 0.8471X2 - 0.7636X3 - 0.0959X4 + 0.1139X5 \\ &+0.2345X6 + 0.4938X7 + 0.1171X8 + 0.0975X9 - 0.0491X10 \\ &+0.0090X11 + 0.1979X12 - 0.0631X13 - 0.5722 \end{aligned} \quad (1)$$

The model demonstrates strong explanatory power ($R = 87.43\%$) and the observed high correlation among variables (0.935) confirms anticipated relationships. The adjusted R^2 is slightly lower (80%) but still significant, indicating a robust and reliable predictive model. The standard deviation is relatively low (0.26) suggesting minimal dispersion. The variables with the greatest weight in the model include prior participation in cybersecurity and cyberdefense simulations ($X2$), prior participation in maritime cybersecurity and cyberdefense simulations ($X3$), time management($X6$), and resources($X7$).

C. Prospective Scenario Analysis

Based on the multiple linear regression model, three prospective scenarios were developed, as shown in Table III. These scenarios serve as reference frameworks, enabling future exercises to be compared against them in order to determine which scenario the observed results most closely align with.

Variable	Pessimistic	Trend-based	Optimistic
Y	2.8	3.8	4.8
X1	1	1	1
X2	0	0	1
X3	0	0	1
X4	0	1	1
X5	3	4	5
X6	3	4	4
X7	3	4	5
X8	3	4	4
X9	3	4	5
X10	2	4	4
X11	1	3	4
X12	2	3	4
X13	1	2	3

Table III: Prospective scenarios considered in our analysis

The **pessimistic scenario** is one where attributes related to i) participants, ii) TTX execution, and iii) presented scenario, are valued with a low or moderate score, having as consequence a low satisfaction ($Y = 2.8$), and a low ability to apply the knowledge acquired along the TTX for the resolution of a CSA situation ($X13 = 1$). The lack of experience and preparation in key areas results in participants struggling to make informed decisions under crisis conditions. Regarding attributes related to participants, this scenario considers situations where participants have scarce experience in general TTXs, and almost null experience or knowledge in cybersecurity simulations, maritime cyberdefense exercises, and cybersecurity and cyberdefense concepts, i.e. $X1 = 1, X2 = 0, X3 = 0, X4 = 0$. Regarding attributes related to TTX execution, this scenario assumes situations where the facilitator has a moderate performance in terms of instructions, time management and delivered resources, i.e. $X5 = 3, X6 = 3, X7 = 3$. Finally, regarding attributes related to the presented scenario, this scenario considers cases where the studied case has deficiencies in terms of relevance, easy-to-understand, complexity, and contributed-details, i.e. $X8 = 3, X9 = 3, X10 = 2, X11 = 1, X12 = 2$.

The **trend-based scenario** is one where attributes related to i) participants, ii) TTX execution, and iii) the presented scenario are valued with medium scores, resulting in a moderate level of satisfaction ($Y = 3.8$) and a medium-low ability to apply the knowledge acquired during the TTX for resolving a CSA situation ($X13 = 2$). This outcome reflects incremental improvements compared to the pessimistic scenario, particularly in aspects related to TTX execution and the scenario itself. While participants benefit from some prior training in cybersecurity and cyberdefense, and the TTX execution is better structured, the scenario complexity remains moderate, limiting the depth of decision-making improvement. These improvements lead to more informed decision-making, but gaps still exist in the application of acquired knowledge during a cyber crisis. The trend-based scenario assumes that participants have some prior experience in crisis simulation exercises or TTX $X1 = 1$, but limited experience in cybersecurity and cyberdefense simulations $X2 = 0$ and maritime-specific cyberdefense simulations $X3 = 0$. Additionally, participants have received some form of formal or informal training in cybersecurity and cyberdefense $X4 = 1$, which provides a foundational understanding of key concepts but lacks practical application in high-pressure scenarios. In terms of TTX execution, the facilitator performs well, providing clear instructions, managing time effectively, and delivering adequate resources, i.e. $X5 = 4, X6 = 4, X7 = 4$. However, the facilitator's performance, while improving compared to the pessimistic scenario. Finally, regarding the presented scenario, the case exhibits moderate relevance and is relatively easy to understand, but the complexity remains at an intermediate level, meaning the scenario presents a reasonable challenge but does not push participants to their maximum decision-making limits. The scenario also provides a moderate level of detail, enough to foster learning but not sufficient for comprehensive, advanced-level strategic decision-making, i.e. $X8 = 4, X9 = 4, X10 = 4, X11 = 3, X12 = 3$.

The **optimistic scenario** is one where attributes related to i) participants, ii) TTX execution, and iii) the presented scenario are valued with high scores, resulting in a high level of satisfaction ($Y = 4.8$) and a high ability to apply the knowledge acquired during the TTX for resolving a CSA situation ($X13 = 3$). This scenario reflects the best-case scenario, where all key aspects align to maximize learning and decision-making capabilities. Participants are well-prepared, the TTX execution is highly effective, and the presented scenario offers the highest level of relevance and challenge. These conditions allow participants to make the most informed decisions, confidently applying their knowledge in resolving complex cyber crises. The optimistic scenario assumes that participants have significant prior experience in crisis simulation exercises or TTX $X1 = 1$, and have participated in cybersecurity and cyberdefense simulations $X2 = 1$, including maritime-specific cyberdefense simulations $X3 = 1$. Additionally, participants have received formal or informal training in cybersecurity and cyberdefense $X4 = 1$, which provides them with advanced knowledge and practical skills. In terms of TTX execution, the facilitator performs excellently, providing clear instructions, managing time effectively, and delivering optimal

resources $X5 = 5, X6 = 4, X7 = 5$. The facilitator's role in providing dynamic, real-time feedback during critical phases ensures that participants are fully engaged and able to learn from the simulation. Finally, regarding the presented scenario, the case exhibits high relevance, is easy to understand, and presents high complexity, challenging participants to apply advanced decision-making skills under pressure. The scenario provides a high level of detail, allowing participants to develop comprehensive, strategic responses to complex cyber crises $X8 = 5, X9 = 4, X10 = 4, X11 = 4, X12 = 4$.

The prospective scenario analysis indicates that the success of TTX exercises and cybersecurity simulations depends significantly on their design and relevance to participants. Strengthening participation in maritime cybersecurity and cyberdefense simulations ($X3$) is crucial, as the difference in satisfaction between the pessimistic and optimistic scenarios suggests that prior simulation experience ($X1$) is a key factor. Ensuring well-trained facilitators with dynamic methodologies can improve participant engagement ($X5$), while realistic and well-structured scenarios enhance immersion ($X8 - X12$). Effective time management ($X6$) and resource ($X7$) allocation are also essential elements that influence participant overall satisfaction (Y), requiring careful planning and execution. Strategic adjustments in these areas can significantly improve the perception and effectiveness of future exercises.

VII. Conclusions

This study introduces a novel hybrid TTX methodology for maritime cybersecurity and cyberdefense, integrating traditional TTX techniques with a mathematical simulation-based model (SERDUX-MARCIM). This approach creates a dynamic and realistic training environment, allowing participants to engage with simulated cyber threats and make informed decisions based on real-time data. By combining serious gaming with simulation modeling, it addresses key limitations of traditional TTX, enhancing both the learning experience and decision-making capabilities.

The methodology was successfully applied in two exercises in Buenos Aires and Washington D.C., using real-world cyber crisis scenarios that demonstrated its effectiveness in strengthening CSA among strategic-level stakeholders in the maritime sector, including military personnel, maritime authorities, and policymakers. These exercises provided valuable insights into improving crisis management and response capabilities. The positive feedback from participants highlights the methodology's potential for broad adoption in maritime cybersecurity.

For future work, it is proposed to increase both the number of participants and the number of exercises conducted, thereby enhancing the statistical robustness and reliability of the results. With larger sample sizes, dimensionality reduction techniques could also be considered to optimize the regression model and focus on the most significant variables. Additionally, the implementation of a pre-exercise assessment is recommended to measure participants' progress throughout the activity. Finally, aligning the evaluated competencies with recognized cybersecurity education and training standards will strengthen the methodological rigor and educational relevance of the proposed TTX.

References

[1] T. Grance, T. Nolan, K. Burke, R. Dudley, G. White, and T. Good, "Guide to test, training, and exercise programs for it plans and capabilities," NIST, Tech. Rep., 2006. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-84.pdf

[2] ENISA, "Good practice guide on national exercises," European Union Agency for Cybersecurity, Technical Report, December 2009. [Online]. Available: https://www.enisa.europa.eu/publications/good-practice-guide-on-national-exercises

[3] ISO, *Societal security — Guidelines for exercises*, International Organization for Standardization Standard ISO 22 398:2013, 2013.

[4] U. Franke and J. Brynielsson, "Cyber situational awareness–a systematic review of the literature," *Computers & security*, vol. 46, pp. 18–31, 2014.

[5] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[6] G. N. Angafor, I. Yevseyeva, and Y. He, "Game-based learning: A review of tabletop exercises for cybersecurity incident response training," *Security and privacy*, vol. 3, no. 6, p. e126, 2020.

[7] V. Švábenskỳ, J. Vykopal, M. Horák, M. Hofbauer, and P. Čeleda, "From paper to platform: Evolution of a novel learning environment for tabletop exercises," in *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, 2024, pp. 213–219.

[8] S. K. Damodaran and N. Wagner, "Modeling and simulation to support cyber defense," pp. 3–4, 2020.

[9] D. J. Bodeau, C. D. Mccollum, and D. B. Fox, "Cyber wargaming: framework for enhancing cyber wargaming with realistic business context," Homeland Security Systems Engineering and Development Institute (HSSEDI), Tech. Rep., 2018. [Online]. Available: http://www.mitre.org/HSSEDI

[10] N. M. Katsantonis, I. Kotini, P. Fouliras, and I. Mavridis, "Conceptual framework for developing cyber security serious games," in *IEEE Global Engineering Education Conference, EDUCON*, vol. April-2019. IEEE Computer Society, 4 2019, pp. 872–881.

[11] J. Valente and M. Reith, "Cyber game-based learning for dod ces," in *European Conference on Cyber Warfare and Security*, vol. 23, 2024, pp. 795–802.

[12] B. Onduto, "Gamification of cyber security awareness – a systematic review of games," Ph.D. dissertation, University of Turku, 12 2021. [Online]. Available: https://www.utupub.fi/bitstream/handle/10024/152929/Onduto_Barack_Thesis_Final.pdf

[13] D. E. Cabuya-Padilla and C. A. Castaneda-Marroquin, "Marco de referencia para el modelamiento y simulación de la ciberdefensa marítima - marcim: estado del arte y metodología," *DYNA*, vol. 91, no. 231, p. 169–179, ene. 2024. [Online]. Available: https://revistas.unal.edu.co/index.php/dyna/article/view/109774

[14] D. Cabuya-Padilla, D. Díaz-López, J. Martínez-Páez, L. Hernández, and C. Castaneda-Marroquin, "Serdux-marcim: Maritime cyberattack simulation using dynamic modeling, compartmental models in epidemiology and agent-based modeling," *International Journal of Information Security*, vol. 24, p. 122, 2025. [Online]. Available: https://doi.org/10.1007/s10207-025-00985-6

[15] T. Grance, T. Nolan, K. Burke, R. Dudley, G. White, and T. Good, "Guide to test, training, and exercise programs for it plans and capabilities," 2006.

[16] N. Abbatemarco, G. Salviotti, L. De Rossi, and C. D'Ignazio, "Understanding leadership competencies in cyber crisis management: Insights from the maersk global supply chain meltdown," in *Proceedings of the 57th Hawaii International Conference on System Sciences*, 2024. [Online]. Available: https://hdl.handle.net/10125/106892

[17] D. T. Wesley, L. A. Dau, and A. Roth, "Cyberattack: The maersk global supply-chain meltdown," *Northeastern University - Ivey Publishing*, 2019. [Online]. Available: https://iveypubs.my.site.com/store/s/product/cyberattack-the-maersk-global-supplychain-meltdown/01t5c00000CwquXAAR

[18] S. Sawang, *Understanding crisis management in modern societies*. Springer International Publishing, 2023, pp. 1–16. [Online]. Available: https://doi.org/10.1007/978-3-031-25188-7_1

Variable	Range	Name	Variable Definition	Guiding Question for Data Collection
Y	0 to 5	Overall Satisfaction	Participants' overall perceived satisfaction with the TTX exercise, considering all aspects of the experience.	Evaluate your overall satisfaction level with the TTX.
X1	0 or 1	Participation in TTX	Participant's prior experience in crisis simulation exercises or Tabletop Exercises (TTX), which may influence their perception of the current exercise.	Have you ever participated in a Tabletop Exercise (TTX) or a crisis simulation exercise?
X2	0 or 1	Participation in CS&CD Simulations	Participant's background in cybersecurity and cyberdefense simulations, which may affect their familiarity with TTX concepts and dynamics.	Have you ever participated in a cybersecurity or cyberdefense simulation exercise?
X3	0 or 1	Participation in Maritime CS&CD Simulations	Participant's prior experience in simulations specifically focused on cybersecurity and cyberdefense in the maritime domain.	Have you ever participated in a cybersecurity or cyberdefense simulation exercise in the maritime domain?
X4	0 or 1	Cybersecurity and Cyberdefense Training	Academic background, training, or informal education received by the participant in cybersecurity and cyberdefense.	Have you received any formal or informal training in cybersecurity or cyberdefense?
X5	0 to 5	Facilitator Effectiveness	Clarity and effectiveness of the instructions, guidance, and assistance provided by facilitators during the TTX.	How clear were the instructions and assistance provided by the facilitators?
X6	0 to 5	Time Management	Assessment of time management during the exercise, including task distribution and the adequacy of time allocated to each phase of the TTX.	How well was time managed during the TTX?
X7	0 to 5	Resources	Quality and adequacy of the materials provided (e.g., game instruction booklet) to facilitate understanding and execution of the exercise.	Did the booklet (game instructions) contain sufficient information for conducting the TTX?
X8	0 to 5	Scenario Relevance	The relevance of the TTX to maritime security awareness and its applicability to participants' operational context.	How relevant was this maritime cybersecurity TTX for raising awareness of maritime security?
X9	0 to 5	Scenario Information Relevance	Value and usefulness of the general scenario information in understanding events and their implications.	How relevant was the general scenario information for understanding the situation and the implications of each event?
X10	0 to 5	Scenario Understanding	Participants' level of comprehension of the presented scenario, considering the information provided for each event of the exercise.	How would you rate your overall understanding of the SCENARIO presented, considering the information provided per event?
X11	0 to 5	Scenario Complexity	Perceived difficulty level of the scenario in terms of exercise structure, involved variables, and presented challenges.	How would you rate the overall complexity of the scenario?
X12	0 to 5	Scenario Detail Level	Accuracy and specificity of the simulated scenario elements, including the fidelity of event and situation representation.	How would you assess the level of detail in the simulated scenario?
X13	0 to 5	Application of knowledge and skills	Applicability of the knowledge and skills acquired during the exercise to the participant's professional responsibilities in maritime cybersecurity.	How applicable are the skills and knowledge acquired in this maritime cybersecurity exercise to your professional responsibilities?
X14	0 to 5	SERDUX-MARCIM Information Relevance	Degree to which the information provided by the SERDUX-MARCIM model was useful and pertinent for understanding the scenario and its implications.	How relevant do you believe the information related to SERDUX-MARCIM was to understand the situation and implications in every event?
X15	0 to 5	SERDUX-MARCIM Information Accuracy	Degree of precision and correctness of the information and visual outputs generated by the SERDUX-MARCIM model.	Considering the scenario presented, how precise do you believe the information and graphics related to the SERDUX-MARCIM were?
X16	0 to 5	SERDUX-MARCIM Need for Similar Models Across Organizations	Perceived importance of implementing simulation models like SERDUX-MARCIM to support decision-making during cyber crisis situations.	How relevant do you believe that implementing cyber threat simulation models (like SERDUX-MARCIM) is for decision-making in a crisis situation?
X17	0 to 5	SERDUX-MARCIM Integration	Perceived level of difficulty in integrating simulation models like SERDUX-MARCIM into current maritime cybersecurity protocols and organizational frameworks.	How complex do you perceive integrating cyber threat simulation models (like SERDUX-MARCIM) into existing maritime cybersecurity protocols and frameworks within your organization?
X18	0 to 5	SERDUX-MARCIM Expected Applicability of the model	Extent to which the SERDUX-MARCIM model's output behavior met the required accuracy to achieve the exercise objective.	Was the output behavior of the SERDUX-MARCIM model sufficiently accurate to meet the exercise objective?
X19	0 to 5	General Observations	Comments or observations related to the exercise development.	N/A

Table II: Definition of Variables and Guiding Questions (Variable Type: numerical)

The QoQo Docker: A Docker Environment for Learning Quantum and Post-Quantum Cryptography in Python

Miguel Salas Heras
Universidad Carlos III de Madrid
Departamento de Informática
miguelsalasheras@gmail.com

Ana I. González-Tablas
Universidad Carlos III de Madrid
Departamento de Informática
aigonzal@inf.uc3m.es

***Abstract*—This article introduces the QoQo Docker, a hands-on environment for learning quantum and post-quantum cryptography in Python deployed on a Docker container. The QoQo Docker includes the pyca/cryptography, Qiskit and liboqs libraries. The QoQo Docker also includes the necessary modules to code and run Python programs and Jupyter notebooks using the Visual Studio Code IDE from the host where the QoQo Docker is deployed. The QoQo Docker is designed for easy deployment and as a lab platform that can be used to support an introductory cryptography course that covers classical (modern) cryptography, quantum cryptography and post-quantum cryptography from a practical perspective. The main target audience is university students with a background in computer science, mathematics, and quantum technologies. An example of the integration of the QoQo Docker into a Master's course focused on quantum and post-quantum cryptography is provided. The results of this integration are satisfactory.**

***Index Terms*—Cryptography, Quantum Cryptography, Post-Quantum Cryptography, Python, Docker, Quantum Key Distribution.**

Contribution type: *Original investigation*

I. Introduction

Quantum and post-quantum cryptography are increasingly vital areas of study within the field of cybersecurity, primarily due to the advent of quantum computing, which has the potential to disrupt traditional cryptographic systems. Quantum cryptography uses the principles of quantum mechanics to enhance security, as exemplified by techniques such as quantum key distribution (QKD), which provides secure communication channels that are resistant to eavesdropping. On the other hand, post-quantum cryptography refers to cryptographic algorithms designed to remain secure against the capabilities of quantum computers, which can theoretically solve hard mathematical problems that underpin current encryption methods, such as RSA and ECC, in significantly shorter timescales. As quantum computers become more powerful, they pose a real threat to these widely-used algorithms, potentially rendering them insecure and exposing sensitive data. Consequently, there is an urgent need to develop and implement post-quantum cryptographic solutions to ensure the integrity and confidentiality of information in a future where quantum computing becomes mainstream [1]. Moreover, post-quantum solutions should also be prepared to integrate QKD and other quantum cryptographic algorithms, as they can provide an additional layer of protection against emerging threats [2].

Navigating this scenario is a major challenge for organisations and industries. It is essential to educate and train individuals with the the necessary knowledge and skills to transition from current cryptographic solutions to robust post-quantum alternatives. Therefore, there is a growing need for educational resources in this area, especially given the high specificity of these cutting-edge areas and the limited availability of learning tools.

In this paper, we present the QoQo Docker, a hands-on environment for learning to code quantum and post-quantum cryptographic algorithms in Python, deployed on a Docker container. The QoQo Docker is designed for easy deployment and as a lab platform that can be used to support an introductory cryptography course covering classical (modern) cryptography, quantum cryptography and post-quantum cryptography from a practical perspective. The main target audience is university students with a background in computer science, mathematics and quantum technologies, although it can also be used to introduce these tools to developers in a practical way.

II. Related Work

Given the critical need to prepare future cybersecurity professionals and empower current ones to deal with the transition of cryptographic tools to post-quantum ones in systems and organizations, there are some previous works that propose educational tools for this purpose.

Focusing on quantum cryptography, López-Incera *et al.* [3] present a game-based approach to teach Bell inequalities and quantum cryptography at undergraduate level. Also with a game-based approach, Escanez-Exposito *et al.* in [4] propose an expansion of *QuBit: The Game* [5] —an innovative board game to learn the concepts behind quantum computing— to introduce quantum key distribution protocols. In [6], as part of the QuVis Quantum Mechanics Visualisation Project, a suite of four interactive simulations that demonstrate the basic principles of three different QKD protocols is developed. Vadla *et al.* [7] present QuaSim, a gamified intelligent tutoring system (ITS) developed to teach quantum cryptography and in which students solve problems related to quantum cryptography through different plans.

Focusing on post-quantum computing, Pérez-Ramos *et al.* [8] propose using the GeoGebra environment to introduce

the basic mathematical concepts of post-quantum cryptography in a visual and very accessible way. There are some publicly available examples of post-quantum cryptography courses (such as the one offered by the Massachusetts Institute of Technology (MIT) in Spring 2024[1], the one offered by the Rochester Institute of Technology (RIT) in Spring 2022, Spring 2023 and Spring 2024[2] [9, 10], or the one offered by the RWTH Aachen University in Winter 2023 and Winter 2024[3]). However, the focus of these courses is more theoretical and mathematical rather than practical hands-on, lacking opportunities for coding or programming using popular languages such as Python or C. For example, the RIT course uses the Sage mathematics software[4] as its primary tool for exercises, which limits engagement with real-world cryptographic libraries.

In addition, while the Open Quantum Safe project [11] provides several Docker containers using different flavours of the liboqs library[5], it does not provide an integrated educational solution as comprehensive as the QoQo Docker. To the best of our knowledge, there is no other publicly available educational Docker that includes all the cryptographic libraries contained in the QoQo Docker, making it a novel and essential resource for both students and professionals in the field.

III. The QoQo Docker Environment

A. Description

Docker[6] is a platform that enables developers to automate the deployment of applications within lightweight and portable containers. These containers encapsulate all the necessary components, including libraries, dependencies, and configuration files, allowing applications to run consistently across different computing environments. Using Docker to set up a coding environment greatly simplifies the process by eliminating the complexity of manual installations and dependency management. Instead of configuring a system from scratch, students can pull a pre-configured Docker image that contains the specific libraries and tools required for their projects. Additionally, Docker provides a coding environment that is reproducible and isolated from other applications.

We have taken advantage of Docker to create an environment where all the necessary libraries and modules are already installed to provide a hands-on experience of coding modern, quantum and post-quantum cryptographic algorithms. The QoQo Docker includes the pyca/cryptography, Qiskit and liboqs libraries. The QoQo Docker also includes the necessary modules to code and run Python programs and Jupyter notebooks using the Visual Studio Code IDE from the host where the QoQo Docker is deployed. Next, we describe these and other supporting components and how they contribute to the QoQo Docker:

- pyca/cryptography[7]. It is a package that provides cryptographic recipes and primitives to Python developers and it is maintained by the Python Cryptography Authority (PyCA). It includes most modern cryptographic primitives and deals with X.509 certificates, allowing a practical experience with the main cryptographic tools used nowadays.
- qiskit-ibm-runtime[8]. Qiskit Runtime is a service built on IBM Cloud that offers the ability to execute quantum workflows on IBM hardware and simulators all within the Qiskit software development kit. Qiskit Runtime ships with Qiskit, which is IBM's open-source quantum software development kit (SDK) built in Python. This library provides a seamless hands-on experience of implementing basic QKD protocols such as the BB84 protocol [12].
- qiskit-aer[9]. Qiskit Aer is high-performance quantum computing simulators with realistic noise models. It provides interfaces to run quantum circuits with or without noise using multiple different simulation methods.
- tno.quantum.communication.qkd_key_rate[10]. TNO Quantum provides generic software components aimed at facilitating the development of quantum applications. The package provides Python code to compute optimal protocol parameters for different quantum key distribution (QKD) protocols [13, 14]. Particularly, it includes modules to support the classical error-correction protocols Cascade and Winnow, that can be used in the error correction phase of a QKD protocol.
- cryptomite[11]. It is a modular, extensible high-level Python library of randomness extractors, created by Quantinuum's Quantum Cryptography. The library provides state-of-the-art randomness extractors that can be used in the privacy amplification phase of a QKD protocol [15].
- liboqs[12]. It is an open-source C library that provides implementations of quantum-resistant cryptographic algorithms. It is designed to prepare cryptographic systems for the era of quantum computing, focusing on algorithms that are secure against potential quantum attacks as a support to the NIST Post-Quantum Cryptography standardization project[13]. It is a critical resource for anyone looking to adopt quantum-resistant cryptography in their systems and applications. liboqs is part of the Open Quantum Safe (OQS) project, that is part of the Linux Foundation's Post-Quantum Cryptography Alliance.
- liboqs-python[14]. It is the OQS's Python wrapper for liboqs. It includes examples to run the well-known algorithms Kyber [16] (for key encapsulation and decapsulation) and Dilithium [17] (for digital signatures).

Additionally, the QoQo Docker includes python (required to run Python code) and the metapackage jupyter (required to run Jupyter notebooks, which is specially interesting for adding visualization support for Qiskit).

[1] https://mitqcrypto.github.io/
[2] https://cs.rit.edu/~tjb/qrc/syllabus_2235.html
[3] https://qis.rwth-aachen.de/teaching/24ws/postquantum-crypto/
[4] https://www.sagemath.org/
[5] https://hub.docker.com/u/openquantumsafe
[6] https://www.docker.com/
[7] https://github.com/pyca/cryptography
[8] https://github.com/Qiskit/qiskit-ibm-runtime
[9] https://github.com/Qiskit/qiskit-aer
[10] https://github.com/TNO-Quantum/communication.qkd_key_rate
[11] https://github.com/CQCL/cryptomite
[12] https://github.com/open-quantum-safe/liboqs
[13] https://csrc.nist.gov/projects/post-quantum-cryptography
[14] https://github.com/open-quantum-safe/liboqs-python

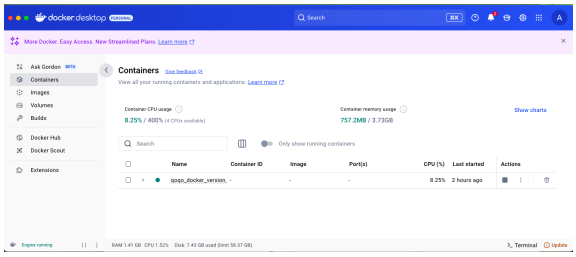


Figure 1. The QoQo Docker running on Docker Desktop

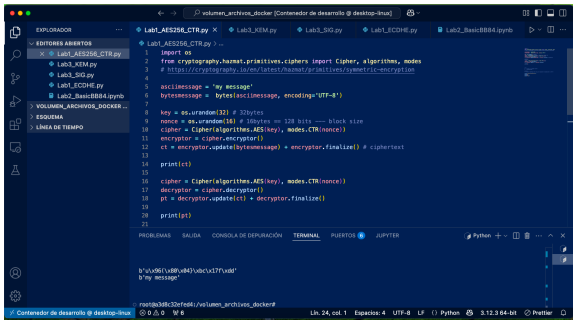


Figure 2. Visual Studio Code running a Python script (using pyca/cryptography) in the QoQo Docker

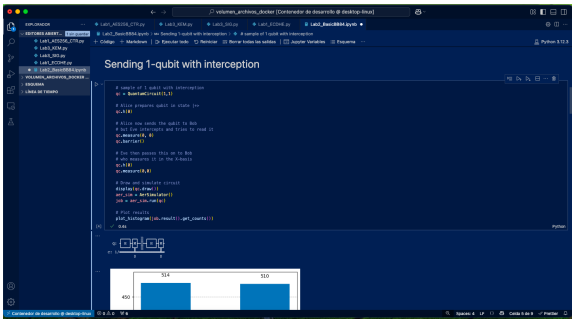


Figure 3. Visual Studio Code running a Jupyter notebook (using Qiskit) in the QoQo Docker

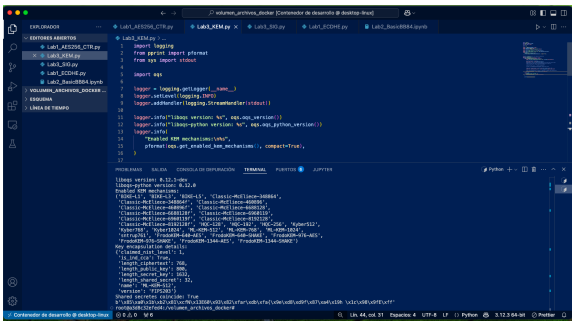


Figure 4. Visual Studio Code running a Python script (using liboqs) in the QoQo Docker

B. Installation and Setup

Detailed instructions on how to install the QoQo Docker and the required files are published in the following public GitHub repository [15]. The following dependencies need to be already installed in the *host* to run the QoQo Docker:

- Docker Desktop[16]: Open source software for deploying applications in isolated software containers.
- Visual Studio Code (VSC)[17]. Integrated Development Environment (IDE) for source code developed by Microsoft.
- Git. Version control software designed by Linus Torvalds. It can be used to clone the repository on the target PC. It is optional.
- Extension dev-containers by Visual Studio Code. This extension allows you to run Docker containers and to replace the execution environment used and displayed in Visual Studio Code with that of the Docker container.

Once the QoQo Docker is deployed, Visual Studio Code can be used to code Python scripts and Jupyter notebooks that run in the QoQo Docker. Figure 1 illustrates how Docker Desktop shows the QoQo Docker running. Figures 2, 3 and 4 illustrate how the QoQo Docker can run code using the libraries pyca/cryptography (a simple AES256-CTR encryption), qiskit (a basic 1-qubit transmission with interception and the visualization of the results), and liboqs (a simple execution of the KEM sample provided with the liboqs-python wrapper).

[15]https://github.com/MichiOnes/QoQo_Quantum_Cryptography_Docker
[16]https://www.docker.com/
[17]https://code.visualstudio.com/

C. Quantifying the Impact of Libraries on the QoQo Docker Image Size

Once the QoQo Docker image has been built, it occupies a considerable amount of space, at approximately 2.8 GB. Table I shows the results of analysing the increase in the image size when adding each of the included libraries and packages. Note that each image builds on the previous one, so the last image in the Table includes all the libraries and packages.

Table I
CUMULATIVE PACKAGE CONTRIBUTIONS TO THE QOQO DOCKER IMAGE SIZE

CONTAINER IMAGE	SIZE
ubuntu	139 MB
ubuntu_with_packages_upgrades	221 MB
ubuntu_with_python_pip_and_git	835 MB
ubuntu_with_pip_upgrade	856 MB
ubuntu_with_liboqs_libraries	884 MB
ubuntu_with_python_numpy_package	992 MB
ubuntu_with_python_liboqs_package	993 MB
ubuntu_with_python_qiskit_aer_package	1,83 GB
ubuntu_with_python_qiskit_ibm_runtime_package	1,94 GB
ubuntu_with_python_tno_quatum_comm ...qkd_key_rate_package	2,26 GB
ubuntu_with_python_cryptography_package	2,7 GB
ubuntu_with_python_jupyter_package	2,7 GB
ubuntu_with_python_cryptomite_package	2,78 GB

The results highlight that the qiskit-aer package and the software required to run Python (Python, pip and git) are the primary contributors to the QoQo Docker image size, accounting for 30% and 22% of the final size respectively. Other important contributors are pyca/cryptography and tno.quantum.communication.qkd_key_rate, which account for the 16% and the 12% respectively. Although minor

adjustments may be possible (e.g., cryptomite somehow overlaps with tno.quantum.communication.qkd_key_rate), these libraries and packages are essential for practical, real-world training in modern, quantum and post-quantum cryptography.

IV. Using the QoQo Docker for Teaching Quantum and Post-Quantum Cryptography

A. Educational Context

The QoQo Docker has been specifically designed to support the practical hands-on sessions of a 6-7 week course on Quantum Cryptography. This course is the second part of the full semester course "Quantum Internet and Quantum Cryptography"[18], which is taught in the Master in Quantum Technologies and Engineering of the Universidad Carlos III de Madrid.

The target students of the Master must have a good background in mathematics and physics, a general knowledge of the technological world, and good fundamentals of programming and algorithms. They belong mainly to two profiles:

- Students with a background in physics (ISCED-F 2013 0533 Physics [18]). The majority of the students enrolled in the academic years 2023-2024 and 2024-2025 fall into this profile.
- Students with a background in electronics and automation (ISCED-F 2013 07143 Electronics and Automation), including the development of electronic equipment, machinery and systems, and the design of computers and equipment for communication[19]. Other suitable backgrounds include specific fields belonging to the broad field of information and communication technologies (ISCED-F 2013 0612 Database and network design administration and ISCED-F 2013 0613 Software and applications development and analysis). Few of the enrolled students fall into this profile.

The Master's program is taught over three semesters. During the first two semesters, students must complete 60 ECTS credits in compulsory courses, and in the third semester, they must take 18 ECTS credits in elective courses plus the Master's Thesis (worth 12 ECTS credits).

As might be expected, most of the courses of the Master focus on quantum technologies and engineering. There are some non-quantum courses (worth 12 ECTS credits) that aim to provide or review basic knowledge in physics (Electromagnetic optics and photonics, and Sensors and classical measurement instrumentation systems), electronics and automation (Pre-quantum information and communication), and information and communication technologies (Secure computer network systems).

B. Course Description

The course has two main groups of objectives as follows:

- Quantum Internet: To understand the fundamental concepts of the Quantum Internet and its distinctions from the classical Internet, explore its relationship with the basics of quantum mechanics, identify the core components of quantum networks –including end systems, communication lines, switches, and routers (repeaters)–and analyse the architecture and implementation of protocols within these networks.
- Quantum Cryptography: To learn the fundamentals and applications of quantum cryptography, especially key distribution, while also understanding the basic concepts, concrete tools and practical deployments of post-quantum cryptography, and to comprehend the impact of quantum computing on systems and infrastructures that incorporate cryptographic algorithms, as well as the challenges of systematically migrating current cryptography to quantum-resistant alternatives

The course is delivered by two different departments. The Quantum Internet part is delivered by the Department of Telematic Engineering, while the Quantum Cryptography is delivered by the Department of Computer Science and Engineering, both belonging to the School of Engineering of the Universidad Carlos III de Madrid. The course runs for 14 weeks, with two 90-minute sessions each week. From now on, the focus will shift to the second half of the course, which covers the Quantum Cryptography part in 15 sessions.

The learning activities include theoretical and practical laboratory sessions, together with individual student work and assessment activities. Table II shows the course schedule, which is divided into four modules addressing the following topics:

1) modern cryptography,
2) quantum key distribution,
3) post-quantum cryptography, and
4) transition to quantum-resistant cryptography and agility.

Each module is associated with a practical laboratory exercise that allows students to enhance their understanding of theoretical concepts and experience the challenges of real-world cryptographic implementation.

Table II
Course Schedule

Ses.	Topic
1	Lec.: Overview of classical (modern) cryptography
2	Lab 1: Alice and Bob communicate securely using modern cryptography
3	Lec.: Quantum Key Distribution (I)
4	Lec.: Quantum Key Distribution (II)
5	Lab 2: Alice and Bob communicate securely using quantum cryptography (I)
6	Lab 2: Alice and Bob communicate securely using quantum cryptography (II)
7	Assessment of Labs 1 & 2
8	Assessment of Labs 1 & 2
9	Lec.: Post-Quantum (Quantum-Resistant) Cryptography (I)
10	Lec.: Post-Quantum (Quantum-Resistant) Cryptography (II)
11	Lab 3: Alice and Bob communicate securely using quantum-resistant cryptography
12	Transition to Quantum-Resistant Cryptography and Cryptographic Agility
13	Lab 4: Alice and Bob communicate securely with everything they have
14	Assessment of Labs 3 & 4
15	Assessment of Labs 3 & 4

The assessment system consists of an oral defence of the laboratory projects (30%), an individual lab test (10%), and

[18]https://aplicaciones.uc3m.es/cpa/generaFicha?est=379&asig=19579&idioma=2

[19]Note that ISCED-F 2013 07143 includes, among others, electronics engineering; automation; computer engineering; communication systems; and telecommunications, network and data processing technologies.

a final test on the theoretical concepts (60%). The lab project is carried out in groups of two students.

C. Labs Description

Lab 1. Alice and Bob communicate securely using modern cryptography. First, students familiarise themselves with working with bytes and files. Then, they explore the use of basic modern cryptographic primitives such as AES256-CTR, SHA256, HMAC-SHA256, HKDF-SHA256 and AES256-GCM. They, then, implement scripts for Alice and Bob to run a Diffie–Hellman key agreement. Finally, they exchange messages securely via files using the agreed key (authenticated encryption with AES256-GCM). For this lab, students are advised to use the pyca/cryptography library and to code using Visual Studio Code.

Lab 2. Alice and Bob communicate securely using quantum cryptography. In this lab, students implement the BB84 quantum key distribution protocol in the presence of noise, utilising the Qiskit and qiskit-aer libraries. Note that other prepare-and-measure or entanglement-based quantum key distribution (QKD) protocols could also be considered, either instead of or in addition to the BB84 protocol. Students analyse the bit error rate in scenarios that combine the presence of an adversary accessing the quantum channel and noise. Messages exchanged through the classical channel are authenticated using HMAC-SHA256 with a pre-shared key. Students also implement information reconciliation using the Cascade or Winnow protocols and privacy amplification using the Trevisan algorithm or a hash function (utilising the tno.quantum.communication.qkd_key_rate and cryptomite libraries). Students are advised to use Jupyter Notebooks in Visual Studio Code for this lab,.

Lab 3. Alice and Bob communicate securely using quantum-resistant cryptography. In this lab, students first explore basic KEM and DSA usage via the examples provided with the liboqs-python wrapper. They then implement a basic key exchange and a double key exchange using ML-KEM or Kyber, as well as an authenticated double key exchange (authentication of at least one-side is mandatory). Authentication of the KEM public keys is achieved using pseudo-certificates issued by a pseudo-Certification Authority that the students implement with ML-DSA or Dilithium signatures. All the keys, ciphertexts and signatures are saved to and loaded from files. Students are advised to code in Visual Studio Code utilising the liboqs-pyhton wrapper and liboqs.

Lab 4. Alice and Bob communicate securely with everything they have. The students are also introduced to hybrid solutions that combine (pre-)quantum and post-quantum cryptography, or quantum and post-quantum cryptography. Students are advised to use the code from previous labs to build the code for this lab.

D. Results

Due to the specific requirements of the cryptographic libraries used in the course, difficulties arose during the practical sessions of the 2023-2024 academic year (the first year of the Master's degree). Students had to install the different libraries and packages, which affected the educational value of the sessions. For this reason, the QoQo Docker was developed and provided to students for use in the hands-on sessions of the 2024–25 course. Consequently, students on the 2024–25 course did not encounter any issues related to installing the libraries, but instead had to install the necessary software to run Docker containers and code in the Python language using Visual Studio Code.

At the end of the practical sessions, students were asked to complete a short anonymous survey assessing their experience of installing and using the QoQo Docker in the course. We received 15 responses from the 20 enrolled students.

Student profile Students use their laptops to complete the practical assignments of the course. Most of them have a Windows operating system (80%). 40% of the students consider themselves to have basic knowledge of installing libraries and software, while 47% think they have reached an intermediate level in this task. Only 13% of the students consider themselves to be very skilled.

Installation Most of the students (66.7%) attended at least one of the installation support sessions (sessions 2 and 5), and the majority of the students (80%) needed additional support during these sessions or via email in addition to the instructions provided in the GitHub repository. The majority of the students needing help were experiencing issues with the DevContainer extension for Visual Studio Code in Windows platforms and a recently updated library version.

Usage and general opinion All the students answering the survey except one have used the QoQo Docker to complete the lab assignments. All the students (100%) find the QoQo Docker a suitable environment for completing the lab assignments.

Opinion on specific functionalities Most of the students (73.3%) think that Visual Studio Code is an excellent tool for writing code, while 13.3% find it good enough, but it took them some time to get used to it. The remaining students find it neither good nor bad, just functional.

Regarding the Jupyter Notebook functionality, students expressed a moderate satisfaction (75% found it very useful and used it, while 25% found it an interesting feature but did not use it). Students' opinions on the shared folder functionality are divided into thirds among "It was very useful (…)" (35.7%), "Useful, but I used it only occasionally" (28.6%) and "I think it is an interesting feature but I didn't use it" (35.7%).

Impact on students' learning Finally, 80% of the students believe that the QoQo Docker has helped them achieve the learning goals of the course, while the remaining 20% feel that it did not make a significant difference.

Instructor perception From the instructors' point of view, once the initial problems with installing the QoQo Docker were resolved, the remaining sessions ran very smoothly, with the students focused on the lab assignments. The lab assessment results were excellent, with students confidently demonstrating their knowledge and understanding of the concepts during their oral defences of the lab assignments and clearly indicating that they enjoyed the practical exercises.

In addition, all the oral defenses of Labs 3 and 4 were carried out on a student's laptop with the QoQo Docker installed. As these final labs use the liboqs library, they benefit most from the tool.

V. Challenges and Limitations

Integrating all the libraries into the QoQo Docker was challenging, but rewarding. The deployment of the QoQo Docker on the laptops of the students enrolled in the 2024-2025 course was satisfactory. Although there were some issues with installing the required dependencies (mainly on Windows platforms), these were resolved during the two support sessions offered to the students. Once the issues were resolved, the QoQo Docker runs well with all the libraries provided in it.

One limitation that has yet to be addressed is the size of the container, which requires about 3GB of space to deploy and run. Currently, the QoQo Docker builds on top of an Ubuntu image and only includes what is absolutely necessary. Other lighter operating systems could be considered, at the cost of more effort to install all the necessary libraries. However, as analysed in Section III-C, the operating system is not the component contributing more prominently to the final size of the QoQo Docker.

Furthermore, the QoQo Docker is not designed to simulate a quantum network (e.g., it does not include software like SimulaQron[20]). This would nicely complete its functionality, but at the cost of requiring much more computational resources.

VI. Conclusion and Future Work

In this work, we have introduced the QoQo Docker, a hands-on environment for learning quantum and post-quantum cryptography in Python deployed on a Docker container. The container and its deployment instructions are publicly available to the community on the GitHub repository https://github.com/MichiOnes/QoQo_Quantum_Cryptography_Docker, to promote education on quantum and post-quantum cryptography, and to foster collaboration with the academic community.

The QoQo Docker has been tested successfully on a small sample of 20 students (those enrolled in the 2024-2025 course). However, some issues arose during the installation that were resolved with the help of support sessions. We expect to continue using it in future editions of the course, which will facilitate further testing and enhancements. Future work will involve publishing a set of curated lab resources to be run on the QoQo Docker.

References

[1] European Comission. *Recommendation on a Coordinated Implementation Roadmap for the transition to Post-Quantum Cryptography*. Tech. rep. Accessed: 2025-03-28. European Union Agency for Cybersecurity, 2023. URL: https://www.enisa.europa.eu/publications/recommendation-on-a-coordinated-implementation-roadmap-for-the-transition-to-post-quantum-cryptography.

[2] Andrea G. Rodríguez. *A quantum cybersecurity agenda for Europe II: Enabling policy and investment options for the quantum transition*. Tech. rep. Accessed: Accessed: 2025-03-28. European Policy Center, 2025. URL: https://www.epc.eu/content/Cybersecurity_2.pdf.

[3] Andrea López-Incera, Andreas Hartmann, and Wolfgang Dür. "Encrypt me! A game-based approach to Bell inequalities and quantum cryptography". In: *European journal of physics* 41.6 (2020), p. 065702. DOI: 10.1088/1361-6404/ab9a67.

[4] Daniel Escanez-Exposito, Pino Caballero-Gil, and Javier Correa-Marichal. "Aprendizaje sobre Computación y Criptografía Cuánticas a través de un Juego de Mesa". In: *Proceedings of RECSI 2024*. León, Spain, 2024. URL: https://hdl.handle.net/10612/24608.

[5] Daniel Escanez-Exposito, Javier Correa-Marichal, and Pino Caballero-Gil. "Qubit: The Game. Teaching Quantum Computing through a Game-Based Approach". In: *Computational Science – ICCS 2023*. Springer Nature Switzerland, 2023. DOI: 10.1007/978-3-031-36030-5_9.

[6] Antje Kohnle and Aluna Rizzoli. "Interactive simulations for quantum key distribution". In: *European Journal of Physics* 38.3 (2017), p. 035403. DOI: 10.1088/1361-6404/aa62c8.

[7] Srikanth Vadla et al. "Quasim: A multi-dimensional quantum cryptography game for cyber security". In: *Journal of The Colloquium for Information Systems Security Education*. Vol. 6. 2. 2019, pp. 19–19. URL: https://cisse.info/journal/index.php/cisse/article/view/101.

[8] Édgar Pérez-Ramos, Pino Caballero-Gil, and Héctor Reboso-Morales. "WIP: Using GeoGebra to Learn the Basics of Post-Quantum Cryptography". In: *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE. 2024, pp. 1–5. DOI: 10.1109/FIE61694.2024.10893100.

[9] Thomas J. Borrelli et al. "Towards a Quantum-Resistant Future: Experiences in Post-Quantum Cryptography Education". In: *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*. Association for Computing Machinery, 2025. ISBN: 9798400705328. DOI: 10.1145/3641555.3705271.

[10] Thomas J Borrelli, Monika Polak, and Stanisław Radziszowski. "Designing and Delivering a Post-Quantum Cryptography Course". In: *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 2024, pp. 137–143. DOI: 10.1145/3626252.3630823.

[11] Douglas Stebila and Michele Mosca. "Post-quantum key exchange for the internet and the open quantum safe project". In: *International Conference on Selected Areas in Cryptography (SAC 2016)*. Vol. 10532. LNCS. Springer. 2017, pp. 14–37. DOI: 10.1007/978-3-319-69453-5_2. URL: https://openquantumsafe.org.

[12] Charles H Bennett and Gilles Brassard. "Quantum cryptography: Public key distribution and coin tossing". In: *Theoretical computer science* 560 (2014), pp. 7–11. DOI: 10.1016/j.tcs.2014.05.025.

[13] Thomas Attema, Joost W Bosman, and Niels MP Neumann. "Optimizing the decoy-state BB84 QKD protocol parameters". In: *Quantum Information Processing* 20.4 (2021), p. 154. DOI: 10.1007/s11128-021-03078-0.

[20] http://www.simulaqron.org/

[14] Xiongfeng Ma, Chi-Hang Fred Fung, and Hoi-Kwong Lo. “Quantum key distribution with entangled photon sources”. In: *Physical Review A—Atomic, Molecular, and Optical Physics* 76.1 (2007), p. 012307. DOI: 10.1103/PhysRevA.76.012307.

[15] Cameron Foreman et al. “Cryptomite: A versatile and user-friendly library of randomness extractors”. In: *Quantum* 9 (2025), p. 1584. DOI: 10.22331/q-2025-01-08-1584.

[16] Roberto Avanzi et al. “CRYSTALS-Kyber - Algorithm Specifications And Supporting Documentation (version 3.02)”. In: *Submission to round 3 of the NIST post-quantum project* (2021). URL: https://pq-crystals.org/kyber/data/kyber-specification-round3-20210804.pdf.

[17] Shi Ba et al. “CRYSTALS-Dilithium – Algorithm Specifications and Supporting Documentation (version 3.1)”. In: *Submission to round 3 of the NIST post-quantum project* (2021). URL: https://pq-crystals.org/dilithium/data/dilithium-specification-round3-20210208.pdf.

[18] UNESCO Institute for Statistics. *Fields of education and training 2013 (ISCED-F 2013) – Detailed field descriptions*. Accessed: 13/05/2025. 2015. DOI: 10.15220/978-92-9189-179-5-en.

ACKNOWLEDGEMENTS

The work of Ana I. González-Tablas is supported by the EU Horizon 2020 Research and Innovation Program under Grant Agreement No. 101021377 (TRUST aWARE), by grant PID2022-140126OB-I00 of the Spanish AEI and by project Vital-IoT funded by INCIBE (Ministry of Digital Transformation and Public Function) and the European Union NextGenerationEU in the framework of the Recovery and Resilience Facility (RRF).

Sesión Investigación VIII: Concienciación y temas emergentes en ciberseguridad

Ciberseguridad y delitos sexuales en víctimas menores de dieciséis años

Dra. María Pilar Marco Francia*
Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Derecho Procesal y Derecho Penal
Facultad de Derecho
Universidad Complutense de Madrid
Pilar.marco@ucm.es

Dr. Juan Carlos Ortiz-Pradillo
Departamento de Derecho Procesal y Derecho Penal
Facultad de Derecho
Universidad Complutense de Madrid
juancarlosortiz@ucm.es

Dra. María Teresa Montero Ruiz
Departamento de Derecho Público y de la Empresa
Facultad de Ciencias Jurídicas y Sociales
Universidad de Castilla-La Mancha
Teresa.montero@uclm.es

Dra. Ana Lucila Saldoval Orozco
Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Ingeniería de Software e Inteligencia Artificial (DISIA)
Universidad Complutense de Madrid
Asandov@ucm.es

***Resumen*- La ciberseguridad es un concepto amplio que comprende la batalla contra un gran número de delitos que son susceptibles de comisión a través de Internet y que no para de crecer todos los días. Las victimizaciones existentes en materia de delitos contra la libertad sexual que sufren menores de dieciséis años, siguiendo esa tendencia, no paran de crecer. Nuestro interés es recalcar la importancia de tener en cuenta las cuestiones criminológicas respecto a las agresiones sexuales que sufren los menores, así como la tipología penal que comprende los delitos susceptibles de comisión a través de Internet, así como la necesidad de conocerlos mejor para poder prevenirlos mediante la implantación de una cultura de la ciberseguridad que impregne todos los niveles, comenzando por la familia y la escuela. Para ello, conoceremos más las características de los delincuentes que abusan sexualmente de menores y sobre todo, de las diferencias de aquéllos que operan online y examinaremos brevemente qué medidas se están adaptando para la localización y enjuiciamiento del autor.**

***Index Terms*- Ciberseguridad, Internet, Derecho Penal, Sexo, Delitos sexuales**

Tipo de contribución: *Investigación original (límite 8 páginas).*

I. CUESTIONES PENALES Y CRIMINOLÓGICIAS RESPECTO A LAS AGRESIONES SEXUALES A MENORES DE DIECISÉIS AÑOS

1. Introducción

En la realidad criminológica, el «abuso sexual» consiste en que, un adulto u otro niño, que puede ser tanto un hombre como una mujer, fuerce o incite a un niño o joven, sin que necesariamente tenga que existir mucha violencia (ya que puede bastar el engaño), a tomar parte en actividades sexuales, sepa el niño o no lo que está pasando. Esas actividades sexuales pueden ir desde: actividades sin contacto físico en las que el niño observe actividades sexuales realizadas por otro o imágenes sexuales, que se involucre en actividades sexuales de forma inapropiada; realizar actos preparatorios para ganarse su confianza, en persona o a través de Internet u otro medio telemático o telefónico, con la finalidad de cometer un abuso sexual de contacto después (se trata del denominado por la doctrina en nuestro país como *on-line grooming*); así como actividades sexuales de contacto, y las cuales pueden consistir en actividades con penetración como, violaciones o práctica de felaciones, o actividades sin penetración como: tocamientos, besos, masturbación. Actividades de carácter no físicas pueden ser: exhibicionismo, pornografía infantil, tanto realizar fotos como obligar al menor a consumirla, u obligar a un menor a participar en actos sexuales con otro menor.

Afecta a niños de ambos sexos y de todas las edades, siendo la más habitual de 7 a 13 años, pero una cuarta parte de los casos denunciados tienen menos de 5 años.

No parece que haya factores de riesgo de raza y estatus socioeconómico, pero sí se incrementan las ratios en los supuestos de padres emocionalmente ausentes por cuestiones psiquiátricas, de dependencias a alcohol o tóxicos, o violentos.

Sí existe un mayor riesgo para menores institucionalizados, que están conformados por menores de estatus económico bajo.

Las denuncias de niñas son de casi el doble con respecto a los niños, aunque los niños victimizados, que tienen una edad mayor que las niñas en el momento del primer abuso, denuncian menos que las niñas.

Existe más probabilidad de que los niños sean victimizados por alguien externo a la familia directa y es más posible que sean abusados por mujeres o por abusadores reincidentes. En el caso de los niños con discapacidad el riesgo de abuso es mucho mayor, y en el caso del abuso sexual se duplica.

2. Algunos datos estadísticos

En la base de datos del Instituto Nacional de Estadística (en adelante INE) se distingue entre condenados mayores de edad y menores. En lo que se refiere a los condenados mayores de edad, el número de condenados se ha multiplicado por cuatro desde 2013 (163) hasta 2023 (839), debido en parte al incremento de la edad de consentimiento sexual de los trece a los 16 años. El detalle sería el siguiente:

	2017	2018	2019	2020	2021	2022	2023
Total	320	453	529	560	738	762	839

Tabla 1: Estadística condenados mayores. Fuente: INE.

Respecto al número de condenas por delitos sexuales realizados por menores contra menores de dieciséis años estos se han incrementado en los últimos años siendo 2022 el año en que más condenas hubo con 385. Las mujeres jóvenes que fueron condenadas por este delito fueron muy escasas.

	2017	2018	2019	2020	2021	2022	2023
Hombres	79	178	241	175	289	385	318
Mujeres	0	3	4	2	2	4	4
Total	79	181	245	177	291	389	314

Tabla 2: Estadística condenados menores de edad. Fuente: INE.

II. LAS AGRESIONES SEXUALES CONTRA MENORES DE DIECISÉIS AÑOS

Tras la modificación de la LO 10/22, de 6 de septiembre, de garantía integral de la libertad sexual (posteriormente modificada por la LO 4/2023, de 27 de abril), el nombre del capítulo se ha modificado pasando a denominarse «De las agresiones sexuales a menores de dieciséis años», conteniendo los artículos 181 a 183 bis del Código Penal. El artículo 181 contiene una panoplia de tipos que examinaremos a continuación:

A) Tipo básico.
El tipo básico se contiene en el artículo 181.1 CP y establece una pena de prisión de dos a seis años, para quien realice actos de carácter sexual con un menor de dieciséis años. El consentimiento en estos supuestos va a ser del todo irrelevante, salvo la cláusula «Romeo y Julieta. Respecto al tipo objetivo: el acto sexual englobará toda aquella acción que posea dicho carácter sexual, incluyendo los que realice el agresor con un menor, el menor con un tercero o incluso si el menor los realiza sobre sí mismo, a instancia del autor. El tipo del artículo 181.1 sería el equivalente de víctimas menores al tipo del artículo 178.1 para los mayores, con la agravación correspondiente a la mayor vulnerabilidad de los menores víctimas.

B) Tipo atenuado.
En el punto tercero del artículo 181 CP se contiene un supuesto atenuado en virtud de la menor entidad del hecho, imaginemos que la agresión sexual es de baja entidad, el órgano sentenciador, puede, valorando las circunstancias y razonándolo en sentencia, puede imponer la pena de prisión inferior en grado. No existirá esa posibilidad si existe violencia o intimidación o se realice sobre una víctima que tenga anulada su voluntad por cualquier causa, o concurran las circunstancias agravantes del artículo 181.5 CP.

C) Tipos agravados.
1) Violencia, intimidación privación de sentido, abuso de la situación mental o anulación de la voluntad. El tipo agravado del artículo 181.2 CP impone una pena de prisión de 5 a 10 años en el caso de que se emplee violencia, intimidación o abuso de una situación de superioridad o de vulnerabilidad de la víctima, de igual forma los que se ejecuten sobre personas que se hallen privadas de sentido o de cuya situación mental se abuse y los que se realicen cuando la víctima tenga anulada por cualquier causa su voluntad. Aunque el artículo remite a las modalidades descritas en el artículo 178.2 y 3 el punto segundo ya contiene la conducta descrita en el punto tercero.
2) Violación.
La pena será de prisión de ocho a doce años si el acto sexual consiste en acceso carnal por vía vaginal, anal o bucal, o en introducción de miembros corporales u objetos por vía vaginal o anal. En el caso de que la violación se realice con violencia, intimidación, abuso de superioridad, de vulnerabilidad, privación de sentido, con la voluntad anulada o abuso mental con la pena de prisión de doce a quince años.
3) Tipos agravados en función de la concurrencia de las circunstancias que conllevan la aplicación de la pena privative de libertad en su mitad superior. Si concurren dos o más se aplicará la pena superior en grado (181.6 CP):
a. Cuando los hechos se cometan por la actuación conjunta de dos o más personas.
b. Cuando la agresión sexual vaya precedida o acompañada de una violencia de extrema gravedad o de actos que revistan un carácter particularmente degradante o vejatorio.
c. Cuando los hechos se cometan contra una persona que se halle en una situación de especial vulnerabilidad por razón de su edad, enfermedad, discapacidad o por cualquier otra circunstancia, y, en todo caso, cuando sea menor de cuatro años.
d. Cuando la víctima sea o haya sido pareja del autor, aun sin convivencia.
e. Cuando, para la ejecución del delito, la persona responsable se hubiera prevalido de una situación o relación de convivencia o de parentesco o de una relación de superioridad con respecto a la víctima.
f. Cuando el responsable haga uso de armas u otros medios igualmente peligrosos, susceptibles de producir la muerte o alguna de las lesiones previstas en los artículos 149 y 150 de este Código, sin perjuicio de lo dispuesto en el artículo 194 bis.
g. Cuando para la comisión de estos hechos la persona responsable haya anulado la voluntad de la víctima suministrándole fármacos, drogas o cualquier otra sustancia natural o química idónea a tal efecto.
h. Cuando la infracción se haya cometido en el seno de una organización o de un grupo criminal que se dedicare a la realización de tales actividades. En caso de que en la descripción de las modalidades típicas previstas en los apartados 1 a 3 de este artículo se hubiera tenido en consideración alguna de las circunstancias anteriores, el conflicto se resolverá conforme a la regla del artículo 8.4 del CP. Cuando exista prevalimiento por parte del culpable de su condición de autoridad, agente de esta o funcionario público, se impondrá, además, la pena de inhabilitación absoluta de seis a doce años (181.7 CP).

D) Participación en actos de naturaleza sexual
El artículo 182 regula como acto típico hacer presenciar, con fines sexuales, a un menor de dieciséis años actos de carácter sexual, aunque el autor no participe en ellos. La pena es de prisión de seis meses a dos años. Hacer visionar pornografía o actos sexuales reales. Este tipo tiene un tipo agravado si los actos que se hace presenciar al menor son constitutivos de un delito contra la libertad sexual (182.2 CP). La pena será de

prisión de uno a tres años.

E) Ciberacoso sexual infantil

El ciberacoso sexual, también denominado por la doctrina como *online sex child grooming*, se basa en las acciones deliberadas por parte del delincuente para establecer una relación y control emocional sobre el menor que preparen el terreno para un abuso sexual ulterior sobre el mismo. Para Swan (2011: 250) [2], el *grooming* es una forma de conseguir acceso a las víctimas y ganar su confianza o tomar ventaja sobre las mismas, generalmente a través de un proceso de aprovechamiento de la vulnerabilidad de los menores, donde el agresor pretende ser un amigo con la finalidad de cometer un abuso sexual posteriormente. El artículo 183 consiste, dentro de la tendencia expansiva del Derecho penal, en el adelanto de la punibilidad penal de actos preparatorios (Agustina et al., 2020: 78). En este caso, se trata de actos preordenados a encontrarse con el menor de dieciséis años para cometer cualquiera de los delitos de los artículos 181 y 189 del C.P.

La conducta típica consiste en contactar, a través de medios, generalmente telemáticos o telefónicos, con un menor de 16 años, con la finalidad de conseguir ganarse la confianza del menor, tener posteriormente un encuentro con él con el propósito de cometer un delito de agresión sexual (artículo 181 C.P.) o realizar la denominada pornografía infantil (que no dejan de ser agresiones sexuales a menores documentadas (artículo 189 C.P.). Las formas de contacto tienen que ser a través de Internet, del teléfono o de cualquier otra tecnología de la información y la comunicación. La finalidad que debe orientar al delincuente es que, en ese ulterior contacto, se cometan los delitos de agresión sexual a menores de dieciséis años o generar material de abuso sexual a dichos menores para fines pornográficos. La pena que lleva aparejada es de prisión de uno a tres años o multa de doce a veinticuatro meses. A esta pena se le sumarán (concurso real) las que le correspondan si se han cometido delitos en el encuentro posterior. Como tipo agravado, el artículo 183.1 *in fine* CP señala penas en su mitad superior si el acercamiento se ha obtenido mediante coacción, intimidación o engaño.

Además, en su punto 2 se castiga con una pena de prisión de seis meses a dos años, una idéntica conducta que vaya dirigida a engañarle para que le aporte material de tipo pornográfico propio o le muestre imágenes pornográficas en las que se represente o aparezca un menor, por lo que el material puede ser también de otro menor. Estas conductas son frecuentes por cuanto los delincuentes se amparan en el anonimato del ciberespacio y se aprovechan de la ingenuidad de los menores, que son ajenos a la posibilidad de que, al otro lado de la pantalla, no sea un joven de su edad, sino alguien mayor, que tiene motivos espurios para relacionarse con ellos. De igual forma, conductas, generalmente voluntarias, como el sexting, tienen una amplia prevalencia. Es necesario concienciar a los menores a través de campañas de prevención o de talleres formativos, de la falta de control que tienen sobre las imágenes de temática sexual que puedan hacer de sí mismos y sobre la que pierden el control una vez han salido de su teléfono. También es necesaria la concienciación sobre sí realmente conocen a la persona a la que se las están reenviando, y de la puesta en peligro de su vida, integridad física o intimidad que estas acciones pueden conllevar.

Es interesante reseñar el estudio de Save the Children (2023) [1] donde se analizaron 33 sentencias y observaron que el perfil de la víctima en online grooming las niñas son un 57,4% y 42,6% son niños. La edad media de todos ellos es 13 años y en un 59% de los casos hay más de una víctima. En lo que se refiere al perfil del agresor era un desconocido en un 47,5 % de los casos y carecían de antecedentes penales en un 95% de los casos. En los casos de que el agresor fuera conocido destaca que un 9,8% del total era un educador o maestro, el 16,4% un monitor o entrenador y un 3,3 un familiar, así como un 23% de los casos simplemente se reseña como cometido por un "conocido", sin distinguir como en los casos restantes. Precisamente, es en este delito, donde el número de agresores desconocidos es importante, la ciberseguridad ha de ser entrenada a todos los niveles, especialmente el educativo en la socialización primaria y en la escuela, siendo esencial que los menores sean dotados de herramientas para detector situaciones de riesgo y para poder enfrentar el supuesto si llegado el caso se produce el evento delictivo.

III. LOS ABUSOS SEXUALES REALIZADOS A TRAVÉS DE INTERNET

1. Introducción

El crecimiento y popularización de Internet en los últimos años se ha venido configurando como un medio en el que numerosos pedófilos utilizan la seguridad y el anonimato que les proporciona la red para localizar niños, entablar una supuesta relación de iguales, simulando tener su edad para abusar de ellos posteriormente. Por no hablar del intensísimo tráfico de la mal llamada pornografía infantil, que únicamente es la constancia gráfica de violencia y explotación sexual de víctimas menores, lamentablemente, tan frecuente en nuestro país. El estudio de estos sujetos que actúan en Internet muestra que reinciden frecuentemente y que se involucran en otros delitos sexuales (Seto y Eke, 2005) [3]. Esto puede ser explicado por la opacidad y secretismo que genera la red, donde es relativamente sencillo esgrimir argucias ocultativas, máxime si la víctima objetivo es un niño y se le añade el *feedback* o retroalimentación que estos sujetos generan entre sí cuando interactúan en la red, (intercambiándose materiales que contienen delitos sexuales contra menores y comentarios) proporcionándose en cierta forma, un respaldo moral, técnico y estratégico por personas que comparten sus intereses sexuales desviados y delictivos, soporte del que carecían antes de la aparición de Internet.

Según Palermo (2008) [4], los delitos sexuales han aumentado, a su juicio, en parte debido al progreso tecnológico y señala que, según las estadísticas, en enero de 2012 existían censados en Estados Unidos, 747.408 delincuentes sexuales, 8.000 más que en el año 2011 (2008: 835). Respecto a las tasas de victimización en delitos sexuales de niños entre 12 y 14 años eran en 1995 de 4‰, y en 2018 de 1.6‰. En lo que respecta a los menores entre 15 y 17 eran en 1995 de 8.9‰, y en 2018 de 3.3‰ (United States Department of Justice, 2020: 1) [5].

Además, se están estudiando las posibles diferencias que puede haber entre un delincuente que utiliza Internet para la comisión de los delitos, o bien el que lo hace directamente. A dicho respecto, es interesante un estudio realizado por Elliot y colaboradores (2013) [6] sobre una muestra de 526 criminales que habían utilizado medios directos de contacto, 459 que habían utilizado Internet y que habían utilizado ambos tipos de contacto y que denominaron como «mixtos». En dicho estudio, realizado a través de autoinforme, utilizando un modelo lineal

general, se puso de manifiesto, —comparando diferentes rasgos psicológicos en lo referente a creencias distorsionadas que apoyaban la comisión de delitos, gestión emocional y respuesta social deseada, entre otros rasgos— que existían diferencias entre los tres grupos. Establecieron como hallazgo que el grupo que contactó personalmente tenía menos nivel de empatía y mayor grado de actitudes prodelictivas e impulsividad que los otros dos grupos. El grupo mixto y el de Internet eran más similares, resultando ser el mixto el más empático y también el que decía sufrir mayor angustia personal, así como el que calibraba mejor los pensamientos y sentimientos de otros.

2. Internet y ciberseguridad en delitos sexuales con víctimas menores de dieciséis años

Internet forma parte indeleble de nuestras vidas. Comunicamos más a través de tecnología online que lo hacemos por persona. En el caso de los jóvenes eso es así, mientras otros de más edad permanecen anclados en lo analógico. Por lo tanto, la accesibilidad de las potenciales víctimas ha aumentado en proporción geométrica, al tener muchos de los menores acceso a Internet a través de los teléfonos móviles. El ciberespacio es un lugar altamente hostil y sin embargo, la confianza impregna a los padres que permiten a los menores su exploración independiente y sometida a múltiples peligros.

En la actualidad, casi la mitad de los delitos, en sentido global, se cometen por Internet. El incremento sustancial de delitos sexuales contra menores online se produjo durante la pandemia (Choi y Lee, 2023) [7].

Sin embargo, la justicia y la sociedad sigue anclada en el mundo «real» e ignorando muchas veces las implicaciones tan potentes que tiene la vida digital, especialmente en el caso de los niños. De la misma manera que se puede hacer daño en la vida real y advertimos a los niños de los potenciales extraños que se aproximan a ellos, si les dejamos un móvil o un ordenador con acceso a internet debemos enseñarles a tomar las decisiones necesarias para que lo puedan hacer seguros, además de ejercer un control parental real de los accesos. Internet facilita, en general, un comportamiento desinhibido, también en lo que se refiere al comportamiento sexual, mostrando lo que generalmente está oculto amparado en el anonimato que comporta. Además, Internet proporciona oportunidades a los delincuentes sexuales de menores para conseguir algo que es previo al abuso: el acceso a un menor. Este acceso no es particularmente fácil en la vida real, al gozar los menores de un mayor control paterno o familiar. Y el control es mayor cuanto menores son. Internet, por lo tanto, da la oportunidad al delincuente de ganarse la confianza del menor a través del proceso de *grooming*. Este proceso de «acicalamiento» o «domesticación», en su aspecto de Derecho penal de actos preparatorios, constituye la actividad criminal consistente en hacerse amigo de un niño con la intención ulterior de persuadir al niño para mantener una relación sexual abusiva. Los ganchos que estos individuos utilizan para mantener un encuentro en el mundo real son: viajes, realizarle fotos profesionales, suscribir un contrato de modelo, jugar al ordenador u otras recompensas que les atraigan (Calder, 2002: 11) [8].

También es cierto que, en otras ocasiones, generalmente cuando la diferencia de edad no es muy grande, las víctimas desarrollan lazos fuertes con sus abusadores o incluso manifiestan estar enamorados por lo que será un reto para desarrollar estrategias preventivas y para que la policía pueda detectarlos. En cuanto al género y edad de los agresores, las mujeres tienen una mayor representatividad en el online grooming de la que tienen en los delitos sexuales contra menores en el mundo real, y también está creciendo el número de delitos sexuales cometidos por niños y jóvenes a través de Internet (McAllinden, 2018: 47).

IV. CARACTERÍSTICAS DE LOS ABUSADORES SEXUALES QUE OPERAN ONLINE

Casi toda la investigación en abusos sexuales sobre menores se ha realizado sobre delincuentes identificados, que pueden ser muy diferentes de los no identificados, ya que pueden ser más impulsivos, que adopten menos precauciones para no ser identificados y que sean más compulsivos sexualmente. De igual forma, los que se controlan mejor, cometen menos delitos online y por lo tanto, reducen el riesgo de ser detectados (Seto, 2013: 137) [10]. Si el agresor tiene gran pericia informática, también aumentará la dificultad para ser detectado. Los delincuentes más cuidadosos evitarán los policías infiltrados en la red, para ello se centrarán más en los sistemas de intercambio de archivos *peer-to-peer* y no en los foros o chats. Davison y Martellozzo (2008: 338) [11] mantienen que los delincuentes sexuales que operan en Internet se pueden clasificar en dos categorías, los que captan menores a través del ciberacoso y los que producen y distribuyen material de abuso sexual a menores. Schmidt (2023) [12] indica que algunos subtipos de agresores que no están interesados en contacto físico fuera de Internet con las víctimas y los denomina los guiados por la fantasía vs. los dirigidos al contacto físico.

Como sostienen Schwartz y Southern (2000) [13], el sexo cibernético se ha convertido en el punto de encuentro de parejas anónimas.

Los autores generalizan sobre los ciberabusadores estableciendo como características generales de los mismos: casados, deprimidos, víctimas de abuso sexual, nivel educativo universitario y usuarios frecuentes de Internet. Los de sexo masculino son de mediana edad, mayores que las mujeres y con una mayor compulsividad sexual. Las mujeres ciberabusadoras, por su parte, son más jóvenes, tienen más trastornos de alimentación como bulimia y tienen más trastorno de estrés postraumático que los hombres, manteniendo muchas similaridades con éstos en parafilias, citas y cambios de pareja.

Al parecer, la policía se enfoca en perseguir más los delitos cuanta menor sea la edad de las víctimas, siendo los niños pequeños mucho más escasos como presas por cuanto a menor edad porque tienen una menor capacidad para navegar solos por páginas de Internet.

Es importante la realización de encuestas anónimas de los delincuentes que operan online y que todavía no han sido detectados. Aunque no podemos olvidarnos de los diferentes sesgos que pueden concurrir, como el de deseabilidad, que la información que nos aporten no sea verificable y que no existan datos longitudinales.

Respecto a los estudios de victimización que se realizan y los delincuentes que se detectan en realidad, la desproporción

es tan grande que posiblemente estemos ante la punta del iceberg (Seto, 2013: 138) [10].

En un interesante estudio de Babchishin y colaboradores (2014) [14] se comparó delincuentes sexuales de menores fuera de Internet con delincuentes de pornografía infantil, y otro grupo en el que había mezcla comparando 30 muestras donde el principal hallazgo fue que las diferencias entre los delincuentes eran principalmente debidas a oportunidades diferenciales por el tipo de delito. Los de modalidades comisivas por Internet (con accesos preferenciales por Internet), y los de contacto, tenían más probabilidad de acceder sexualmente a menores que los que únicamente accedían a pornografía infantil. Por su parte, los consumidores de pornografía infantil puntuaban más bajo en valores antisociales y a su vez, mostraban más empatía con la víctima, existiendo menos pedófilos que en los otros dos grupos.

V. LOCALIZACIÓN E IDENTIFICACIÓN DEL AUTOR

A la hora de localizar e identificar al autor de unos hechos cometidos a través de Internet (por ejemplo, difusión de material Tipologías contextuales de abusos y agresiones sexuales a menores pedófilo…), una herramienta indispensable es acceder a la dirección IP y demás datos de tráfico del equipo desde el que se difundió dicho material, para lo cual las autoridades cuentan con las obligaciones establecidas sobre los proveedores de acceso a Internet recogidas en la Ley de Conservación de Datos (Ley 25/2007) y en la Ley de Enjuiciamiento Criminal (Ortiz Pradillo, 2020) [15].

También resulta técnicamente posible llegar a determinar la identidad y localización de un usuario de una red social que haya difundido dicho material pedófilo, a través de herramientas OSINT *(open source intelligence* and social media investigation's techniques) de geolocalización (Ortiz Pradillo, 2021) [16]. En la actualidad, existen también análisis del lenguaje natural a través de programas informáticos, para detectar *a priori* si se utiliza contenido de abuso sexual de menores (Povedano et al., 2023a) [17], a través de algoritmos de inteligencia artificial (Martínez, Sandoval y García, 2023; Povedano et al., 2023b) [18] [19]. Aunque estos análisis están plenamente incorporados al uso forense, también podrían ser muy importantes, por el gran aporte de evidencia psicológica en términos de evaluación de riesgo de reincidencia que se podrían realizar (Drouin et al., 2017: 4) [20].

Aunque del Informe sobre la cibercriminalidad en España de 2023 [21], del total de los delitos denunciados (472.125) únicamente 1804 fueron delitos sexuales (excluidas las agresiones sexuales con o sin penetración y los abusos sexuales con penetración), lo cual implica únicamente un 0,4 % de la cibercriminalidad en España, eso no quiere decir que los hechos sean de gran gravedad y sean susceptibles de provocar graves secuelas en los niños y niñas víctimas, por lo que la prevención a través de la ciberseguridad debe ser una prioridad.

VI. CONCLUSIONES

En materia de delitos sexuales contra menores de dieciséis años, casi todos los supuestos que hemos examinado son susceptibles de comisión en remoto a través de Internet. Como ya hemos señalado, el ciberespacio dista mucho de ser un lugar seguro, en especial sobre menores que vagan por los recursos de Internet sin acompañamiento o supervisión adulta. es un lugar altamente hostil y, sin embargo, la confianza impregna a los padres que permiten a los menores su exploración independiente y sometida a múltiples peligros. Debemos partir de una educación en ciberseguridad con los menores ya que el conocimiento les empodera y libera de riesgos y pensar un modelo en el que, por definición, no se confíe en absoluto en Internet, a fin de mantenerles libres de depredadores sexuales digitales que andan al acecho a fin de conseguir su satisfacción lúbrica digital y posiblemente, en el plano físico después, atentando contra el menor tras haber ganado su confianza, o tras coaccionarle o amenazarle.

AGRADECIMIENTOS

El presente trabajo se ha realizado con los fondos del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation), a través de la Cátedra "Ciberseguridad para la Innovación y la Protección Digital" INCIBE-UCM.

Además, este trabajo ha contado con el apoyo de los siguientes Proyectos de Investigación:

Proyectos de la Comisión Europea: HEROES (acuerdo de subvención nº. 101021801) y ALUNA (https://aluna-isf.eu/), en virtud del acuerdo de subvención nº 101084929. El contenido de este artículo no refleja la opinión oficial de la Unión Europea ni de la Comisión Europea. Los autores son los únicos responsables de la información y las opiniones expresadas en él.

Proyecto «Violencia extrema sobre la mujer: respuesta penal y alternativas para la prevención» (VEM, PID2022-140170NB-I00). 2022-2026. CO-IP's Carolina Bolea Bardon y Juan Carlos Hortal Ibarra. Entidad financiadora: Ministerio de Ciencia e Innovación del Gobierno de España.

Proyecto Criminalidad organizada transnacional y empresas multinacionales ante las vulneraciones a los derechos humanos. Nº: PID2020-117403RB-100. IP Laura Zúñiga Rodríguez. Programa financiador: Programa Estatal de I+D+I orientada a los Retos de la Sociedad, del Plan Estatal de Investigación Científica y Técnica y de Innovación de 2017-2020. Entidad financiadora: Ministerio de Ciencia e Innovación del Gobierno de España.

"Cuestiones controvertidas en relación con el ejercicio de la capacidad jurídica por los y las menores de edad" (Ref. CIGE/2023/71), financiado por la Conselleria de Educación, Universidades y Empleo de la Generalitat de Valencia. I.P. P. Chaparro Matamoros."

REFERENCIAS

[1] Save the Children. (2023). Online grooming. Análisis de sentencias sobre online grooming a niños y niñas en España.

[2] Swann, R. (2011). New technologies, new challenges: assessment and treatment guidance for young people's problematic/abusive sexual behaviour online. En: Calder, M. C. (ed.). *Contemporary practice with young people who sexually abuse*. Evidence-based developments, pp. 227-262. Dorset: Russell House Publishing.

[3] Seto, M., Eke, A. (2005). The criminal histories and later offending of child pornography offenders. *Sexual Abuse: A Journal of Research and Treatment*, Vol. 17(No. 2), 201-210.

[4] Palermo, G. B. (2008). An old problem seeking a new solution: Sexual Offenders. *International Journal of Offender Therapy and Comparative Criminology*, (1)(56), 3-5.

[5] United States Department of Justice. (2020). Juvenile justice bulletin. Disponible en: https://ojjdp.ojp.gov/juvenile-violent-victimization.pdf.

[6] Elliot, I., Beech, A., Mandeville-Norden, R. (2013). The Psychological Profiles of Internet, Contact, and Mixed Internet/Contact sex Offenders. *Sexual Abuse: A Journal of Research and Treatmen*t, 1(25), 3-20.

[7] Choi, K-S., Lee, H. (2023). The Trend of Online Child Sexual Abuse and Exploitations: A Profile of Online Sexual Offenders and Criminal Justice Response, *Journal of Child Sexual Abuse*.

[8] Calder, M. C., (2002). *Young people who sexually abuse. Building the evidence base for your practice*. Dorset: Russell House Publishing.

[9] McAlinden, A. M., (2018). *Children as 'risk'. Sexual exploitation and Abuse by children and young people.* Cambridge: Cambridge University Press.

[10] Seto, M. (2013) *Internet sex offenders*. Washington: American Psychological Association.

[11] Davison, J., Martellozzo, E. (2008). Protecting children online: towards a safer internet, en: Letherby, G., Williams, K., Birch, P. y Cain, M., *Sex as a Crime*, (338-355), Devon, Willan Publishing.

[12] Schmidt, A. F. (2023). Sexual Online Grooming of Children. A Short Overview of the Offender Literature. En Kuhle, L. F. & Stelzmann, D. (eds.), *Sexual Online Grooming of Children*. Nomos.

[13] Schwartz M. F. y Southern, S. (2000) Compulsive Cybersex: The New Tea Room, *Sexual Addiction y Compulsivity,* 7:1-2, 127-144.

[14] Babchishin, K. M., Nunes, K. L., & Kessous, N. (2013). A Multimodal Examination of Sexual Interest in Children: A Comparison of Sex Offenders and Nonsex Offenders. *Sexual Abuse*, 26(4), 343-374.

[15] Ortiz Pradillo, J. C. (2020) Europa: auge y caída de las investigaciones penales basadas en la conservación de datos de comunicaciones electrónicas, *Revista General de Derecho Procesal*, núm. 52.

[16] Ortiz Pradillo, J. C. (2021) Big data, vigilancias policiales y geolocalización: nuevas dimensiones de los derechos fundamentales en el proceso penal, *Diario La Ley*, número 9955.

[17] Povedano Álvarez, D., Sandoval Orozco, A. L., Portela García-Miguel, J., García Villalba, L. J. (2023a). Learning strategies for sensitive content detection. *Electronics* 12, n. 11: 2496.

[18] Martínez Hernández, L. A., Sandoval Orozco, A. L., García Villalba, L. J. (2023). Analysis of digital information in storage devices using supervised and unsupervised natural language processing techniques. *Future Internet*, vol. 15, issue 5.

[19] Povedano Álvarez, D., Sandoval Orozco, A. L., García Villalba, L. J. (2023b). Detección de contenido sexual explícito mediante técnicas de aprendizaje profundo. Actas de las VIII *Jornadas Nacionales e Investigación en Ciberseguridad*; Universidad de Vigo.

[20] Drouin, M., Boyd, R. L., Greidanus Romaneli, M. (2017). Predicting Recidivism Among Internet Child Sex Sting Offenders Using Psychological Language Analysis. *Cyberpsychology, behavior, and social networking*, vol 00, number 00.

[21] Dirección General de Coordinación y Estudios. Secretaría de Estado de Seguridad. (2024) Informe sobre la criminalidad en España 2023. https://www.interior.gob.es/opencms/pdf/archivos-y-documentacion/documentacion-y-publicaciones/publicaciones-descargables/publicaciones-periodicas/informe-sobre-la-cibercriminalidad-en-Espana/Informe-Cibercriminalidad_2023_126200212_pdfWEB.pdf

Exploring the impact of Spanish-specific trained PassGPT model for password guessing

Germán Pinedo Terrados
Universidad Carlos III de Madrid
Departamento de Informática
germanpinedoterrados@gmail.com

Ana I. González-Tablas
Universidad Carlos III de Madrid
Departamento de Informática
aigonzal@inf.uc3m.es

***Abstract*—Passwords serve as a primary authentication factor in most websites, applications, and online services. Despite their widespread use, they remain vulnerable, particularly when they are simple and easy to guess. In recent years, artificial intelligence (AI) has increasingly been applied to password guessing, leading to the development of models that can generate highly realistic password lists based on real-world data. It is well established that demographic factors influence password selection, yet no prior research has specifically analysed password guessing performance on Spanish users. In this study, we investigate whether training an AI model, in particular PassGPT, on Spanish-specific passwords improves its effectiveness in guessing passwords compared to a globally trained model. To assess this impact, we conducted a comparative analysis, evaluating the number of successfully guessed passwords for Spanish users. Two models were used: one trained exclusively on Spanish passwords and another trained on a globally sourced dataset without country-based filtering. Our results indicate that language-specific training data significantly affects password guessing performance.**

The Spanish-trained model achieved an accuracy of 12.91%, outperforming the globally trained model, which achieved 9.49%. These findings are aligned with previous research and confirm the importance of considering demographic factors in AI-driven password security research.

***Index Terms*—Password guessing, Deep learning, Language-specific**

Contribution type: *Original investigation*

I. Introduction

Passwords are one of the most common authentication methods for humans due to its simplicity. They are also used as fallback system if other authentication methods fail. Therefore, it is essential for users to have secure, complex and not easy to guess passwords. Despite this, attacks to them are common and according to the 2024 Data Breach Report by the Identity Theft Resource Center [1], the number of accounts compromised (victim notices) during 2024 was more than 1.3B, with a remarkable growth respect to previous years thanks to 5 mega-breaches occurring during year 2024.

While there are a variety of attack vectors that can lead to account compromise, the impact of using compromised credentials is always significant. According to IBM's 2024 Cost of Data Breach Report [2], attackers used compromised credentials in 16% of the breaches studied (with a cost of USD 4.81M per breach), closely followed by 15% of breaches where the attack vector was phishing (with a cost of USD 4.88 million per breach). Likewise, according to Verizon's 2024 Data Breach Investigations Report [3], in 77% of the basic web application attacks related to data breaches[1], attackers gain access by using stolen credentials, and 41% of successful logins across websites protected by Cloudflare involve compromised passwords (on Cloudflare's observed traffic between September - November 2024) [4].

Cybercriminals can compromise credentials using several types of attacks. Although password cracking involves attempting to recover plaintext passwords when password hashes are available, both password cracking and password guessing (which assumes no prior knowledge of the legitimate password) may use repetitive and systematic methods based on dictionaries and lists of common passwords [5]. These traditional password brute force methods are not efficient if the target has complex and highly personalised passwords. Attackers can then take advantage of their knowledge of typical password complexity policies and common password pattern composition practices, including variations that exploit some information about the target (like name, address, birthday...). Tools like John the Ripper and Hashcat are widely used for performing password cracking efficiently, using these techniques and supporting a variety of attack modes [6, 7].

Unfortunately for cybersecurity professionals or daily users, the recent breakthrough of artificial intelligence (AI) can also be harnessed for application in this matter. Several technologies based on AI have been developed for generating lists of passwords (that can be used for password guessing or password cracking), being PassGAN [8] and PassGPT [9] the most prominent ones.

Both have been extensively used by researchers, obtaining remarkable performance results against past password leaks. It should be noted that the aim of these investigations is not to compete with the cybercriminals, but rather to assess the extent to which these techniques can be used and to estimate the risk posed to users and organisations. These studies are generally based on password leaks that are published on the internet, which, depending on the leak, contain information of users from different countries, services...

The passwords that users select are dependant, between other facts, on their country. The culture and language influence the individuals, as they may use words, expressions, places that have relation to their place of origin or residence [10]. Previous research has explored the impact of using language-specific deep learning techniques to improve the success rate of password guessing models (see, for example,

[1]Other typical attacks, not classified as web application attacks, are malware or exploit vulnerabilities.

the work of [11] that leverages this information to improve the performance of PassGAN on an Italian password leak).

The objective of this work is to analyse if the adaptation of the training of an AI model to a specific language, in this case Spanish, increases the guessing results in respect to a more generic approach. Particularly, we have based our work on the recently published PassGPT model, that uses OpenAI GPT-2. To the best of our knowledge, this is the first work exploring Spanish-specific training with a modern deep-learning technique. The scope of the investigation is focused in obtaining a proof of concept rather than developing a very advanced and optimized model that outperforms the results obtained in other investigations, as we are limited by computing capacity and existing data, as will be explained in Section V. To accomplish the objective, we have trained PassGPT from scratch with data that has been properly filtered to ensure that it belongs to Spanish population. Then, a comparison has been done with the results obtained by a more generic approach of the same model, trained with passwords belonging to the whole world. Our results confirm that carrying out a specific training in terms of language, in our case Spanish, has a higher success rate in terms of guessing, as expected from the results of previous work.

The paper is structured as follows. First, in Section III we discuss all the work previously carried out in this line of research. Then, in Section IV, the methodology is explained, from the selection of the dataset to the whole process of training, testing and evaluation. In Section V, the limitations of this study are discussed and, finally, Section VI concludes our work.

II. Background

Cybercriminals utilise various types of password attacks to compromise user credentials. Attackers continue to exploit services and systems by taking advantage of default, weak, or easily guessable credentials. This can be done through brute-force methods—such as password guessing or cracking—as well as by purchasing stolen credentials from illegal marketplaces, thereby skipping the technical effort of obtaining them manually. Another common tactic is reusing credentials leaked in previous data breaches, often through techniques like password spraying, which involves trying common passwords across many accounts, or credential stuffing, where previously leaked username-password pairs are tested in bulk against multiple services.

AI and, specifically, generative AI, greatly improve the effectiveness of previous types of attacks. AI can automate the brute force approach by rapidly generating and testing large numbers of password combinations. In addition, AI can be used to analyse patterns in leaked password databases and create improved password guessing or cracking models.

The two main deep learning password guessing models proposed in the last years are PassGAN [8] and PassGPT [9]. PassGAN is based on Generative Adversarial Networks (GANs), a type of deep learning model where two neural networks, the generator and the discriminator, compete against each other to improve password generation [8]. The generator creates password candidates that imitate real-world passwords, while the discriminator evaluates their authenticity. Over time, this adversarial training process allows PassGAN to learn realistic password patterns and structures from a dataset.

In contrast, PassGPT is built on Generative Pretrained Transformers (GPT), a type of deep learning model designed for generating sequences of text [9]. Unlike GANs, which rely on two competing networks to improve their results, transformers use a different approach called self-attention. This allows them to analyse and understand patterns in large amounts of text, including passwords, before generating new ones. By learning from a vast dataset of real-world passwords, PassGPT can recognize common structures, variations, and patterns, having the advantage to create a wider variety of realistic password candidates.

Since transformers process entire sequences at once rather than refining their output step by step, they can recognize long-term relationships between characters and adapt better to different password styles. GANs are more limited and create less varied passwords and do not take into account the context about passwords, it is less flexible.

III. Related Work

Password-based attacks are a research area that receives significant attention from the cybersecurity community, as password-based authentication has been one of the first authenticator methods used in computer systems and networks and is still the most widely used, either alone or in combination with other methods that increase its robustness (e.g., multi-factor authentication).

Password guessing is one of the easiest way for an attacker to gain access to an account. [12] presents a very nice introduction and historical survey on this specific attack. In addition, Bonneau [12] proposes several metrics to estimate the difficulty of guessing passwords from a certain password distribution and privacy-respectful password cracking techniques based on dictionaries. Results on an anonymized corpus of 70M passwords show that the number of bits of security for different identifiable subgroups of users are very similar. However, they observe a moderate increase in the success rate of password guessing when the dictionary used is matched to the user's preferred language.

In recent decades, the risk posed by password-guessing techniques has increased with the availability of large leaks of passwords or password hashes (e.g., COMB dataset [13]), and with the advent of AI-based techniques (e.g., PassGAN [8] and PassGPT [9]). Later, in Section IV-C, we will give more details on each of these deep learning-based password guessing models.

Most of the research in this area is focused on enhancing the performance of password guessing techniques, but there are some works that focuses on analysing the impact of different languages, cultures or demographics on the user's choice of password in order to exploit these differences to increase the success rate of a particular password guessing technique [10]. Han *et al.* [14] analyse such differences between the passwords of English-dominant users and Chinese users, achieving an increase by up to 48% in the guessing efficiency of their PCFG-based method after including Chinese Pinyins.

Some recent works have explored the use of language-enriched Probabilistic Context-Free Grammars (PCFG) to improve the guesses of passwords. For example, [15] focus

their study in three main European languages (English, French and Portuguese) —achieving over a 10% increase in the success rate with respect to state-of-the-art methods— and [16] propose an improved PCFG-based password-guessing method that adds grammars regarding Asian-language input procedures and pronunciation rules —achieving improvements ranging from 2% to 14% for different password datasets—.

Additionally, Capuani, in his Master Thesis [11], explores the impact of using PassGAN in password cracking for the password leak Libero.it, increasing the success rate from 71% (achieved by state-of-the-art tools) to 80%. More recently, Pasquini *et al.* present a "universal" password model based on Attention Mechanisms that, once pre-trained, exploits user's auxiliary information to adapt the guessing strategy. The empirical results of their proposal on 6 language-specific password distributions (English, Russian, Korean, Polish, French and Taiwanese) confirm the benefits of exploiting this information for password guessing.

Although there is some work on the impact that language-specific training has in password guessing and cracking, as far as our knowledge concerns and having analysed the mentioned related work, there are no investigations regarding Spanish passwords.

IV. Methodology and Results

Throughout this Section, the whole process of the study until getting the results will be explained in depth.

A. Overview

The methodology followed has the main aspects of any AI model training but with adaptations to our computation capacity and dataset. In Figure 1 the different parts of the methodology can be seen. It begins with the *data collection and preprocessing*, in which we select the dataset and preprocess its contents to obtain the different datasets that will be used in the research. As the computational limitations are important, the process of cleaning and filtering had to be developed different for Spanish and World versions because of the high volume of records. After cleaning, filtering and preprocessing, the Spanish dataset is split in two subsets, one to train (TRAIN-SP) and the other to test (TEST-SP). From the World dataset, a subset of records is randomly extracted (TRAIN-WO). This dataset has the same size of TRAIN-SP and will be used to train a World-based model for comparison with the Spanish trained model. Notice that TEST-SP is the only dataset that will be used for any performance evaluation, either it is a Spanish trained model or World trained.

Then, the *model training and hyperparameter tuning* started. First, a trial-and–error approach is applied to figure out how the model behaves. Once better results are found out, a last and finer tuning of the hyperparameters has been done to find the best performance possible.

Finally, in the *model evaluation*, the models that performed best in the training phase are used to generate a big amount of passwords and performance results (number of guessed passwords) are obtained against the testing dataset (TEST-SP). The results are compared with the performance of the same models (configuration, hyperparameters,...) but trained with the World dataset (TRAIN-WO).

B. Dataset selection and preprocessing

In order to train the model, a high volume of data is needed, which is obtained from the COMB dataset, a combination of several password leaks from different sources [13].

COMB is a very heterogeneous dataset, which contains passwords from well-known password leaks, such as "rockyou" or "exploit.in", linked to an email account. It contains approximately 3.28 thousand million passwords, 2.18 thousand million unique emails [17]. It aggregates passwords from different years but also countries, email or service providers, that is the reason why it is interesting to use this leak to train a password guessing model. Previous research analyses the COMB dataset and some of the leaks it contains, resulting in some descriptive statistics [18]. As COMB is a compilation of breaches, the data contained in it may have little variances depending on the year or the source.

Table I
Records of exploit.in data leak by country or purpose

Country / Purpose	Number of records (M)
USA / Commercial	389
Russia	203
Germany	47
France	26
Spain	2.4

Table II
Distribution of records by password length

Password length	Number of records	Password length	Number of records
0	5,429	9	274,633,133
1	4,980	10	64,442,280
2	2,038	15	4,911,355
3	14,644	20	1,731,936
4	194,429	25	334,040
5	16,462,822	30	45,651
6	29,792,114	35	6,582
7	165,972,861	40	1,311
8	56,302,118	45	1,260
–	–	50	364

To have a high-level overview of the characteristics of the dataset, we extract some of the statistics about the exploit.in leak from the cited source. The exploit.in data leak is included in the COMB and can show representative data for COMB. In Table I the number of passwords by country is shown. In addition, the number of emails with a determined password length is indicated in table II. As a detail, it can be seen that some accounts have empty passwords, which can be due to poor secured websites.

In exploit.in data leak, the most popular domain is the United States (or commercial) followed by important some European countries. Also, the most common password length is 9 characters, followed by passwords with a length of 7. There are very few extremely long passwords, just as there are few very short passwords.

COMB has a high volume of passwords and an important amount of them are not useful for this study, so a preprocessing has been applied. The first action was to remove all

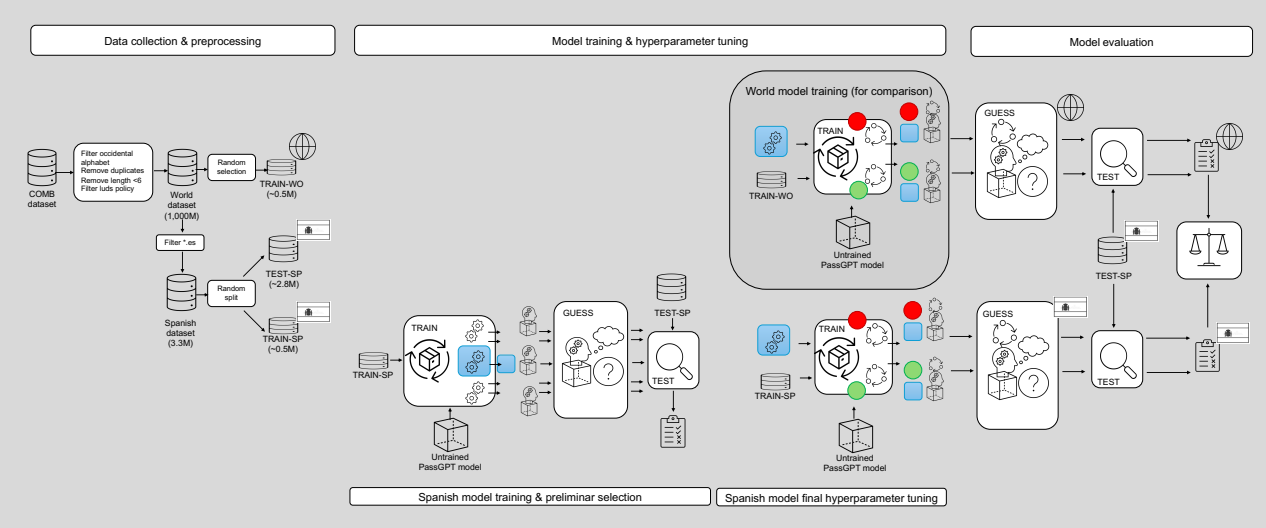


Figure 1. Overview of the methodology

duplicated passwords and the ones that do not use occidental alphabets, for example, removing passwords written with Cyrillic or Chinese characters if there were.

The decision of removing duplicated passwords was a debate. Duplicated passwords may indicate trends in population, so it can serve as a ponderation, giving more importance to those more repeated. As a consequence, the model would theoretically give more importance to the more frequent passwords. This was rapidly discarded due to the computing limitations, but it might be interesting to look into it further.

It is a common practice by digital platforms to implement additional conditions to the passwords of their clients, forcing them to be more complex and consequently increasing the difficulty of guessing attacks. A brief research on some of the most important platforms was carried out to have a criteria on how to discriminate the passwords of the leak. Furthermore, a common use of passwords cracked by criminals is to access those services looking for personal information or the possibility of transferring money, so seems reasonable to take into account the goals of the criminals to further filter the records in the dataset. The results are shown in Table III.

Table III
PASSWORD LENGTH AND COMPOSITION REQUIREMENTS OF SOME MAIN PLATFORMS

Website	Minimum character length	Number of different categories of characters
gmail.com	8	N/A
amazon.com	6	N/A
instagram.com	6	2
outlook.com	6	2
apple.com	8	3

Therefore, as part of the preprocessing, only passwords consisting of at least 6 characters and including at least two of the following three character categories are taken into account: uppercase, lowercase and digits. This procedure is undertaken to ensure that the passwords have a certain degree of complexity and are realistic.

In addition, the preprocessing done makes sense to the objective of creating complex passwords. Basic passwords can be guessed using less sophisticated techniques such as dictionary or rule-based attacks, using open source tools such as John the Ripper or Hashcat [6, 7]. The resulting dataset can be regarded as a realistic collection of human passwords.

To differentiate Spanish records, the domain ".es" of the associated email has been used as a filter. Despite knowing that there are more Spanish records in the dataset, particularly those linked to the ".com" domain belonging to Spanish citizens, it is not possible to differentiate between Spanish and non-Spanish records. Once the domain has been used to filter the records, the email addresses have been deleted, anonymising the records of the dataset.

Note that due to RAM limitations, the scripts used to preprocess the records without filtering by domain (to obtain the world-based dataset) had to use a specific strategy capable of overcoming such limitations.

The final amount of Spanish passwords after the preprocessing is approximately 3.3 million, in contrast with the world-based dataset, whose quantity is around 1,000 million passwords (309 times bigger). The Spanish dataset is divided into two subsets, training (TRAIN-SP) and test (TEST-SP). The final choice for the size of these datasets[2] is 500,000 passwords for training (selected randomly) and the remaining ones for testing (∼2.7 million passwords). The training subset represents almost the 16% of the amount of passwords, while the test subset is the remaining 84% approximately. These proportion is a balance between a good amount of training passwords and enough passwords to reliably test the models. From the world-based dataset, also 500,000 passwords have been randomly extracted to build TRAIN-WO, the training dataset that is used in the evaluation phase to compare the results of a world-based trained model with the Spanish-specific trained one.

[2]Results from preliminary trainings with less passwords brought poorer results.

C. Experimental setup and selection of password generation model

The study is strongly conditioned by the equipment used in the training phase, a laptop with 8GB of RAM, a GeForce GTX 1650 with 4GB of VRAM as GPU and an AMD Ryzen 5 3550H as CPU.

Initially, the two main deep learning password guessing models proposed in the last years were considered as the starting point for this research, i.e., PassGAN [8] and PassGPT [9]. When created PassGPT, a comparison was carried out by the author against PassGAN and the results are quite interesting. PassGPT clearly outperforms PassGAN and the number of unique passwords generated by the model is also greater (see Table 3 and Figure 3 of [9]). As it has been said, PassGPT has a more recent technology, which is rapidly evolving. The version used is OpenAI's GPT-2.0[3] but more recent models exist so PassGPT can be improved if desired [19].

For the previous reasons, PassGPT was the technique chosen for this study. PassGPT allows acceleration using GPU, which allows us to compensate in part the lack of computing resources. The unique modification done to the training source code was in relation to the GPU usage. The model was moved to CUDA in the training script. For this reason, the project was developed using a Conda virtual Python environment, as there are some requirements that require specific software versions in terms of compatibility between CUDA and the different Python modules used.

D. Model training and hyperparameters tuning

The training phase involved multiple iterations with different settings to refine the model's performance. In all the training processes described next, PassGPT was trained from scratch, using a specific training dataset, either TRAIN-SP or TRAIN-WO. Model refinement was done by identifying the best model configuration and hyperparameters in terms of number of guessed passwords of the test dataset (TEST-SP), when the model was used to predict passwords.

The model development process began with an exploratory phase, where various configurations and dataset sizes of guessed passwords (4M and 8M) were tested to determine their impact on password generation accuracy.

Broad configurations were experimented with, including different learning rates, batch sizes, and tokenization methods. A trial-and-error approach allowed for the identification of promising configurations that give better results. The tokenizer selected is the biggest one, with 99 characters including, uppercase, lowercase, numbers and special characters. Using more reduced tokenizer, with only lowercase, numbers and special characters, lead to a slightly faster training but, the result of the models had a decrease in accuracy of approximately 0.5%.

The evaluation of each model was based on its capability to predict passwords contained within the test dataset. The model with the highest number of guessed passwords and, taking into account that the number of guessable passwords is the same, also the one with the highest success rate will be considered the best. There were an approximate maximum of 2.75M guessable passwords for both Spanish and world-based models, as the comparison is always against the Spanish password testing dataset (TEST-SP).

Once the different models are trained, two files containing 4 and 8 million passwords were generated for the evaluation. This was undertaken to evaluate the models performance and determine their relative evolution in proportion to the increase in passwords generated, ensuring that models that behave good on low numbers of generation, maintain their performance on higher generation.

The preliminary training and testing phase, resulted in the selection of Model 3, as its performance is the best one among the selected preliminary models (see Table IV).

Table IV
RESULTS OF BEST SPANISH-TRAINED MODELS DURING PRELIMINARY TRAINING PHASE

Model	Guessing results for 4M		Guessing results for 8M	
	% (of TEST-SP)	N	% (of TEST-SP)	N
Model 1	2.82029	78,766	4.41393	123,274
Model 2	4.18879	116,986	6.40155	178,785
Model 3	**4.58050**	**127,926**	**6.96779**	**194,599**
Model 4	4.27515	119,398	6.50904	181,787
Model 5	4.35568	121,647	6.67228	186,346

Configuration parameters of Model 3 were used as the basis for additional trainings by adding little adjustments, with more powerful configurations, seeking a last performance increase. These enhancements resulted in Model 3+ and Model 3*, which were selected for the evaluation phase.

E. Model evaluation

. In first place, Models 3+ and 3* were tested against the dataset TESP-SP by generating 4M passwords. The outcome of this step can be appreciated in Table V. Results show that the configuration of Model 3+ is the most accurate for the study. Although the difference in performance percentage may seem small, the difference in the total number of passwords guessed is significant, which can be appreciated in Table V.

Table V
RESULTS OF BEST SPANISH-TRAINED MODELS (GUESSING 4M PASSWORDS)

Model	Guessing results for 4M	
	% (of TEST-SP)	N
Model 3+	**4.69601**	**131,152**
Model 3*	4.64198	129,643

Lastly, we compare the results of Models 3+ and 3* trained with the Spanish dataset (TRAIN-SP) with the results of the same models but trained with the world-based dataset (TRAIN-WO). Note that for the comparison with the world-based model, the configurations used for training were the same regardless of the language specificity of the dataset used for training. Table VI shows the final results for the study regarding the efficiency of training a model in the specific language against a generic training.

The results indicate that the specialised training yields superior results in comparison to the generic model, outperforming it by a 3.4%, almost 1.35 times the results, which is a significant difference.

[3] https://huggingface.co/docs/transformers/model_doc/gpt2

Table VI
RESULTS OF BEST SPANISH-TRAINED MODELS AND COMPARISON WITH WORLD-TRAINED MODELS (GUESSING 25M PASSWORDS)

Model	Results for Spanish		Results for the World	
	% (of TEST-SP)	N	% (of TEST-SP)	N
Model 3+	**12.90766**	**360,490**	**9.49464**	**265,170**
Model 3*	12.86358	359,259	9.46278	264,280

F. Model training and evaluation times

Regarding the time required for the different phases, the training phase took around 11 hours for both models (Models 3+ and 3*), while the generation took almost 50 hours for the first and 53 hours for the second. In terms of time, no difference was appreciated between Spanish and world datasets, which is logical considering the amount of passwords is the same.

V. LIMITATIONS AND DISCUSSIONS

Once the results have been presented, an analysis must be made and conclusions drawn, which should take into account the objectives and context of the investigation.

A. Limitations given by computing environment

The study is bounded by the computing capacity of the device used. A more efficient training could have been done but neither the computer had powerful computing resources nor was available the time desired. This conditions forced us to seek a different solution for the whole process.

All changes of parameters have been made by human agency, making the process less efficient, but necessary to run the scripts when it was possible. The common and automatized way of training is to test a big amount of hyperparameters, defining a selection of them with a range of values and trying every combination to then extract the best configuration. This process with the code provided would need a big adaptation, but the big problem is the time it would cost, which made it impossible.

Related with these constraints, we have left for future work several issues such as applying cross validation, more demanding trainings and generating larger password datasets with the models. The latter would directly impact the number of guessed passwords, as there would be more possible successful cases.

B. Limitations given by the size of the Spanish password dataset

The number of available Spanish passwords has also limited the study. 3.3 million passwords is a little amount for a generative model, considering that the world dataset multiplies the number by more than 300. More passwords would have lead to more examples for the model to adapt and then to make a more accurate generation, but, in addition, more guessable passwords may have been available, so the results would have changed radically.

It was also considered to use passwords from Spanish speaking Latin America countries but was discarded due to, although being similar in language and some other aspects, cultures are different in terms of expressions, names, traditions... Those aspects influence the passwords that citizens choose, so the decision to have a highly specialized Spanish model was taken.

The amount of passwords for the world model is limited by the size of the Spanish one, in order to have the same quantity. This can be regarded as unfair for the world, as one of its advantages is to have a larger available dataset, but if the objective is to directly compare its efficiency against a Spanish model, it makes more sense to use the same amount and check how far the results are from it. It would be interesting to study if increasing the number of passwords of the world dataset enhances the accuracy of the model against Spanish passwords, but that is out of the scope of this research.

C. Difficulties on comparing results

Comparing the accuracy of this study with respect to others is a difficult task, as there are some aspects to take into account. The very first is that none or little work has been carried out for specific languages in PassGPT. For PassGAN, we found a study that worked specifically on Italian [11]. It concludes that the language on the training does have an impact.

The dataset used for the Spanish model is in a different language than those used in other studies, which are typically not tailored to any specific country or language. As a result, comparing the Spanish model directly to these studies would be unfair. Instead, the appropriate comparison should be made against a general-purpose, globally trained model.

In addition, if there were more testing passwords, the final number of guessed passwords with the same model and generation number would be higher, as more successful cases exist. In terms of guessing percentage, this cannot be ensured. Other studies use the whole world dataset, with more passwords to train and guess, so the performance is harsh to compare.

Regarding computing capacity, many studies do not give information about the specifications of the computers, hours taken... It has a direct impact on the performance. If no specifications are given, it is difficult to figure out if the model trained is efficient or not in respect of others. With more computing capacity, more passwords could be generated in the same time and the results of the guessing would increase.

D. Relevance of results

Despite the inherent limitations of this study, we believe that the results obtained are highly relevant and meaningful. The entire research was conducted on a personal laptop, equipped with modest hardware specifications and constrained by limited computational time. This context highlights a critical point: it is possible for an individual, without access to advanced infrastructure or specialized knowledge, to generate and crack millions of passwords using publicly available AI models and datasets.

While the results presented here were achieved under constrained conditions, an attacker with greater computational power, more time, and larger datasets could potentially cause even more severe damage, compromising a vast number of user accounts.

VI. Conclusions

Based on the findings, we can conclude that training an AI model on language and country-specific data such as passwords from the Spanish population leads to improved guessing performance compared to a more generic, globally trained model. It is important to emphasize that the goal of this research was not to develop the most efficient or general-purpose model, but to examine the impact of localized training by comparing two different approaches under identical conditions: one tailored to Spanish data and another based on global, language-agnostic data.

The findings obtained throughout the study confirm the significant implications for cybersecurity of the use of modern generative AI models, such as PassGPT, for password guessing. Moreover, these results emphasize the urgent need to reassess current password policies and defence mechanisms. If a single researcher can achieve such outcomes on a limited setup, the risk posed by well-funded malicious actors is considerably greater. This underlines the necessity of promoting stronger authentication mechanisms, such as multi-factor authentication or security keys, the usage of password managers and encouraging users to adopt complex, preferably randomly-created, and unique passwords that are less susceptible to AI-based guessing techniques.

Acknowledgements

The work of Ana I. González-Tablas work is supported by the Spanish National Cybersecurity Institute (INCIBE) grant APAMciber within the framework of the Recovery, Transformation and Resilience Plan funds, financed by the European Union (Next Generation) and by grant PID2023-150310OB-I00 of the Spanish AEI.

References

[1] Identity Theft Resource Center. *2024 Data Breach Report*. Available at: https://www.idtheftcenter.org/wp-content/uploads/2025/02/ITRC_2024DataBreachReport.pdf. Accessed: 18/03/2025. 2024.

[2] IBM. *Cost of a Data Breach Report 2024*. Available at: https://www.ibm.com/downloads/documents/us-en/107a02e94948f4ec. Accessed: 18/03/2025. 2024.

[3] Verizon. *2024 Data Breach Investigations Report*. Available at: https://www.verizon.com/business/resources/reports/dbir/. Accessed: 18/03/2025. 2024.

[4] Zejnilovic Rajwan. *Password reuse is rampant: nearly half of observed user logins are compromised.* Available at: https://blog.cloudflare.com/password-reuse-rampant-half-user-logins-compromised/. Accessed: 21/03/2025. 2025.

[5] MITRE Corporation. *MITRE ATT&CK*. Available at: https://attack.mitre.org/. Accessed: 18/03/2025. 2024.

[6] Rasheed Agboluaje and Olushola Odejobi. "Comparative Analysis of John the Ripper, Cain and Abel, and Hashcat for Password Recovery in Encrypted Files and Archives". Available at: https://www.researchgate.net/profile/Rasheed-Agboluaje-2/publication/388327569_Comparative_Analysis_of_John_the_Ripper_Cain_and_Abel_and_Hashcat_for_Password_Recovery_in_Encrypted_Files_and_Archives/. Accessed: 11/05/2025. 2024.

[7] Kaden Marchetti and Paul Bodily. "John the Ripper: An examination and analysis of the popular hash cracking algorithm". In: *2022 Intermountain Engineering, Technology and Computing (IETC)*. IEEE. 2022, pp. 1–6. DOI: 10.1109/IETC54973.2022.9796671.

[8] Briland Hitaj et al. "Passgan: A deep learning approach for password guessing". In: *Applied Cryptography and Network Security (ACNS 2019)*. Vol. LNCS 11464. Springer. 2019, pp. 217–237. DOI: 10.1007/978-3-030-21568-2_11.

[9] Javier Rando, Fernando Perez-Cruz, and Briland Hitaj. "PassGPT: Password Modeling and (Guided) Generation with Large Language Models". In: *Computer Security – ESORICS 2023*. Springer Nature Switzerland, 2024, pp. 164–183. ISBN: 978-3-031-51482-1. DOI: 10.1007/978-3-031-51482-1_9.

[10] Mashael AlSabah, Gabriele Oligeri, and Ryan Riley. "Your culture is in your password: An analysis of a demographically-diverse password dataset". In: *Computers & security* 77 (2018), pp. 427–441. DOI: 10.1016/j.cose.2018.03.014.

[11] Eugenio Maria Capuani. *Exploring the impact of language differences on GANs for Password Cracking*. Master's thesis. Available at: https://rucforsk.ruc.dk/ws/portalfiles/portal/64898094/thesis_final.pdf. Roskilde University, 2019.

[12] Joseph Bonneau. "The science of guessing: analyzing an anonymized corpus of 70 million passwords". In: *2012 IEEE symposium on security and privacy*. IEEE. 2012, pp. 538–552. DOI: 10.1109/SP.2012.49.

[13] Chloe Stejskal et al. "Analyzing the Impact and Implications of COMB: A Comprehensive Study of 3 Billion Breached Credentials". In: *2024 IEEE 24th International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*. IEEE. 2024, pp. 158–167. DOI: 10.1109/CCGridW63211.2024.00026.

[14] Weili Han et al. "Regional patterns and vulnerability analysis of chinese web passwords". In: *IEEE Transactions on Information Forensics and Security* 11.2 (2015), pp. 258–272. DOI: 10.1109/TIFS.2015.2490620.

[15] Carlos Ocanto Dávila, Rocío Cabrera Lozoya, and Slim Trabelsi. "Sociocultural Influences for Password Definition: An AI-based Study". In: *Proceedings of the 7th International Conference on Information Systems Security and Privacy (ICISSP 2021)*. 2021, pp. 542–549. DOI: 10.5220/0010269305420549.

[16] Shin-Hung Yan and Cooper Cheng-Yuan Ku. "Using language-specific input methods and pronunciation rules to improve the guesses of passwords". In: *Journal of Information Security and Applications* 77 (2023), p. 103588. DOI: 10.1016/j.jisa.2023.103588.

[17] Syhunt Icy Team Daragon. *COMB: The Big Password Leak*. Available at: https://www.syhunt.com/en/index.php?n=Articles.COMBPasswordLeak2021. Accessed: 31/03/2025. 2021.

[18] *Exploit.in Data Breach Analysis*. Available at: https://breachdirectory.com/blog/exploit-in-data-breach-analysis/. Accessed: 21/03/2025. 2020.

[19] Staphord Bengesi et al. "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers". In: *IEEE Access* (2024). DOI: 10.1109/ACCESS.2024.3397775.

(Work-in-progress) Measuring Data Pollution in Cryptocurrency Abuse Reporting Services

Gibran Gomez
IMDEA Software Institute
Universidad Politécnica de Madrid
gibran.gomez@imdea.org

Kevin van Liebergen
IMDEA Software Institute
Universidad Politécnica de Madrid
kevin.liebergen@imdea.org

Davide Sanvito
NEC Laboratories Europe
davide.sanvito@neclab.eu

Giuseppe Siracusano
NEC Laboratories Europe
giuseppe.siracusano@neclab.eu

Roberto Gonzalez
NEC Laboratories Europe
roberto.gonzalez@neclab.eu

Juan Caballero
IMDEA Software Institute
juan.caballero@imdea.org

***Abstract*—Cryptocurrency abuse reporting services collect victim reports and are a valuable data source about abusive blockchain addresses, prevalent types of cryptocurrency abuse, and their financial impact on victims. However, they may suffer data pollution due to their crowd-sourced nature. This work-in-progress analyzes the extent and impact of data pollution in cryptocurrency abuse reporting services. For this analysis, we collect 289K abuse reports submitted over 6 years to the BitcoinAbuse and Scam Tracker services. We show that, if left unchecked, spam reports will eventually flood an abuse reporting service, with BitcoinAbuse receiving 75% of spam before stopping operations. We build a public dataset of 19,443 abuse report textual descriptions labeled with 19 popular abuse types and use it to reveal the inaccuracy of user-reported abuse types. Finally, we identify 91 (0.1%) benign addresses reported, responsible for 60% of all the received funds. Using this dataset we are designing automated techniques for the classification of abuse reports.**

***Index Terms*—abuse reporting services, bitcoin, cryptocurrency, abuse**

Type of contribution: *Work-in-progress*

I. Introduction

Cryptocurrencies have become a popular payment and financial investment mechanism. Unfortunately, they are also frequently abused by cybercriminals, for example, to receive payments from victims of criminal activities such as ransomware [1], [2], [3], [4], [5], [6], sextortion [7], [8], and scams [9], [10], [11], [12], [13]. A valuable source of data on cryptocurrency abuse is reporting services that provide online forms for users to fill out with information about the abuse they have observed (e.g., date, description, type, amount lost), the abuser (e.g., email address, Bitcoin address), and the reporters themselves (e.g., country, age). While some abuse reporting services specialize in cryptocurrency abuse (e.g., BitcoinAbuse [14], ChainAbuse [15]), others collect all abuse reports, regardless of whether they involve cryptocurrencies, such as the Better Business Bureau (BBB) Scam Tracker [16], the Australian Competition and Consumer Commission (ACCC) ScamWatch [17], and the Federal Trade Commission (FTC) ReportFraud [18].

Abuse reports provide critical information on abusive Bitcoin addresses, prevalent cryptocurrency abuse types, and financial losses suffered by victims. Blockchain explorers [19], [20] link Bitcoin addresses to their abuse reports in BitcoinAbuse so that users can avoid known malicious addresses. Unfortunately, abuse reporting services can suffer data pollution due to their crowd-sourced nature. A fraction of the submitted reports may be useless, including advertisements, trolling, reports aiming to pollute the database, and reports not describing abuse. Some services manually validate the reports to address data pollution. Unfortunately, manual validation is costly and does not scale. Thus, a critical challenge is how to identify valid reports automatically.

Abuse reports are also fundamental for analyzing the prevalence and financial impact of different abuse types (e.g., sextortion [7], investment scams [9]). BBB and FTC publish well-known annual reports with statistics on the received abuse reports [21], [22], which prior work often cites in their motivation [23], [24], [25]. For example, in 2023, the FTC received 47K abuse reports involving cryptocurrency payments, totaling $1.4 billion in losses [21]. Those statistics rely on an accurate classification of the abuse reports into abuse types.

This work analyzes the challenge of data pollution in cryptocurrency abuse reporting services. To enable our analysis, we collect cryptocurrency abuse reports from two services. We obtained 287K abuse reports for 92K Bitcoin addresses received by BitcoinAbuse [14] over 6 years. We also collected 2,321 cryptocurrency abuse reports received by Scam Tracker [16] over 5 years.

We use this data to analyze the extent and impact of pollution in abuse reporting services. We examine three types of pollution: spam reports, incorrect abuse types reported by users, and benign addresses reported as malicious. We focuse on BitcoinAbuse, where there was no filtering for submitted reports. Then, we check if the identified issues also affect Scam Tracker. We identify two prevalent types of spam reports in BitcoinAbuse: advertisements for funds recovery services and reports that aim to discredit cryptocurrency exchanges by claiming they are involved in terrorism. At least 10.6% of all BitcoinAbuse reports are spam, with the ratio reaching 75% of submitted reports prior to the service stopping operation. In contrast, Scam Tracker manually validates submitted reports. Still, we observe a handful (0.4%) of funds recovery spam

Dataset	Reports	Desc.	Addr.	Start	End
BitcoinAbuse	287,823	224,059	92,151	03/2017	06/2023
Scam Tracker	2,321	2,311	96	02/2017	05/2024

Table I
ABUSE REPORT DATASETS USED.

reports missed by their filtering.

To measure the accuracy of user-reported abuse types, we built a dataset of 19,443 abuse report descriptions labeled with 19 popular abuse types. We compare this ground truth dataset with the 6 abuse types that reporters could select in the online form (which we call *BA types*). Reports labeled with three BA types (Bitcoin-Tumbler, Darknet-Market, Ransomware) were almost entirely spam. Two other BA types (Blackmail-Scam, Other) were too broad. The only accurate BA type was Sextortion. Even after removing spam reports, users often selected an incorrect abuse type, with 22% of Ransomware-labeled reports being indeed sextortion. In Scam Tracker, analysts classify the reports. However, all cryptocurrency abuse reports are placed in the same class, preventing the analysis of fine-grained abuse types.

Out of the 92,151 Bitcoin addresses reported to BitcoinAbuse, we identify 91 (0.1%) benign addresses, most often addresses of exchanges. 72% of reports for benign addresses are spam, but we also observe users trying to track the destination of their funds and ending up incorrectly reporting an exchange address. Even if benign addresses are a minority, their impact is overwhelming, with those 91 addresses being responsible for 60% of all received funds. While Scam Tracker does not have a specific field for reporting cryptocurrency addresses, the 96 Bitcoin addresses found in the text of reports are not known to be benign.

Previous works that analyzed BitcoinAbuse [26], [27], [28] did not filter out spam reports, considering that all reported addresses were malicious (and thus that their deposits corresponded to cybercriminal revenue) and that the reporter-selected abuse types were correct. Our results cast a shadow over their conclusions.

II. ABUSE REPORTING SERVICES

As summarized in Table I, we collect abuse reports from two services: BitcoinAbuse and Scam Tracker.

a) BitcoinAbuse: Each report contains the report creation date and four fields from the online form: the reported Bitcoin address, the *abuse type* selected from a drop-down list with 6 options, an *other-abuse* short string optionally used to describe other types of abuse, and the description text for freely describing the abuse (2,000 characters maximum). To preserve anonymity, the reports do not identify the reporter.

The dataset comprises 287,823 reports for 92,151 Bitcoin addresses, an average of 3.1 reports per address. The 287,823 reports have 224,059 unique descriptions (by SHA256 of the description text). Multiple reports may have the same description, e.g., due to short descriptions such as "Sextortion" or "Scam". We identify the description's language using the *langdetect* tool [29]. Most descriptions are written in English (80.5%), followed by Russian (1.7%), French (1.7%), Spanish (1.6%), and German (1.6%).

b) Scam Tracker: We build a crawler to download abuse reports from the Scam Tracker website. Scam Tracker groups all reports related to cryptocurrencies in the same *Cryptocurrency* abuse type. In May 2024, we downloaded all 2,321 Cryptocurrency reports. For each report, we obtain the scam ID, report date, description, victim's location, and dollars lost (optional). The 2,321 reports have 2,311 descriptions written in English. The online form does not collect cryptocurrency addresses, but we use the iocsearcher tool [30] to extract 96 Bitcoin addresses present in 71 reports.

III. ABUSE TYPE TAXONOMY AND GROUND TRUTH

We build a taxonomy of prevalent abuse types and ground truth (GT) datasets of descriptions annotated with their abuse type in the taxonomy. Section III-A describes the design of our taxonomy, and Section III-B details the process for generating the ground truth.

A. Taxonomy

Prior work has presented cryptocurrency abuse taxonomies [31], [32], [12]. Our taxonomy builds on those, but it is more complete and specifically designed hierarchically to better accommodate possible incompleteness and facilitate extensions. Figure 1 shows our tree-structured abuse taxonomy, where nodes are abuse types, and edges capture parent-child relationships, e.g., *investment* is a subtype of *scam*. The lower a node, the more specific abuse it captures.

Any abuse taxonomy is necessarily incomplete as cybercriminals frequently introduce new abuse types and variations. Our hierarchical taxonomy is designed to handle such missing abuse types. For example, any scam report that does not match the 7 L3 scam subtypes can be labeled using the L2 *scam* type, indicating that it captures another type of scam. Examining the reports assigned an L2 type (or the L1 *abuse* type) may reveal new abuse types that should be added to the taxonomy.

B. Ground Truth Construction

Since we cannot manually examine the 224,059 BitcoinAbuse descriptions, we design a two-step approach to build our taxonomy in Figure 1 and the GT datasets in Table II. First, we manually analyze 175 descriptions to produce an initial GT (gt_initial) with its corresponding initial taxonomy. Then, we cluster the 224K descriptions by the similarity of their text and manually label a fraction of the clusters. The cluster analysis identifies additional prevalent abuse types not in gt_initial, which we add to the taxonomy. It also produces a larger GT (gt_final) with 19,443 descriptions.

a) Initial ground truth and taxonomy: We built an initial taxonomy by manually labeling 175 descriptions. To select these descriptions, we split the 287K reports in the BitcoinAbuse dataset per year of the report and the six BA types, producing 35 buckets: one per year (from 2018 to 2023) and abuse type (except for Sextortion in 2018 since it was created in 2019). Then, we randomly selected 5 unique descriptions from each bucket, avoiding duplicates. Two analysts manually classified each of the 175 descriptions by reading their text and freely assigning labels to the descriptions based on their experience. After completing the first labeling round, the analysts discussed the selected labels, normalized their names, and merged duplicates, coming up with the 16 abuse types

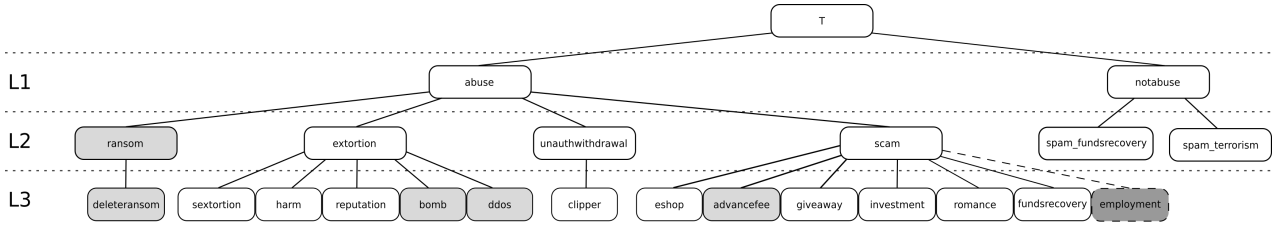


Figure 1. Our taxonomy. White nodes come from the initial taxonomy. Light gray solid nodes were identified by the clustering. Dark gray dashed nodes were not originally part of the taxonomy but were added during our evaluation.

Type	gt_initial	gt_final	gt_bbb
sextortion	44 (25.1%)	16,112 (82.8%)	23 (11.5%)
extortion	28 (16.0%)	1,417 (7.3%)	2 (1.0%)
giveaway	6 (3.4%)	740 (3.8%)	6 (3.0%)
spam_fundsrecovery	1 (0.6%)	174 (0.9%)	-
harm	1 (0.6%)	170 (0.9%)	-
deleteransom	-	162 (0.8%)	1 (0.5%)
investment	11 (6.3%)	143 (0.7%)	102 (51.0%)
advancefee	-	74 (0.4%)	8 (4.0%)
notabuse	27 (15.4%)	73 (0.4%)	8 (4.0%)
spam_terrorism	3 (1.7%)	64 (0.3%)	-
ddos	-	62 (0.3%)	-
scam	31 (17.7%)	60 (0.3%)	48 (24.0%)
reputation	1 (0.6%)	54 (0.3%)	-
bomb	-	52 (0.3%)	-
romance	2 (1.1%)	34 (0.2%)	-
clipper	1 (0.6%)	33 (0.2%)	-
unauthwithdrawal	7 (4.0%)	7 (0.0%)	2 (1.0%)
eshop	6 (3.4%)	6 (0.0%)	-
abuse	5 (2.8%)	5 (0.0%)	-
fundsrecovery	1 (0.6%)	1 (0.0%)	-
ransom	-	-	-
All	175 (100%)	19,443 (100%)	200 (100%)

Table II
DESCRIPTIONS FOR EACH ABUSE TYPE IN GT DATASETS.

in white in Figure 1. Thus, despite the small number of descriptions in gt_initial, this analysis identifies 16 out of the 22 entries in the final taxonomy. Then, the analysts performed a second joint labeling round to assign one of the 16 identified abuse types to the 175 descriptions.

b) Final taxonomy and ground truth: Figure 1 shows the final taxonomy, where the light gray nodes correspond to the 5 abuse types introduced by the cluster analysis. Table II describes the final ground truth (gt_final), which contains all gt_initial descriptions, plus those in the clusters manually labeled by the analyst. In total, gt_final contains 19,443 descriptions. The top abuse type is *sextortion* with 16,112 (82.8%) descriptions, followed by *extortion* (7.3%), *giveaway* (3.8%), and *spam_fundsrecovery* (0.9%).

c) Scam Tracker ground truth: We apply the same process for generating gt_initial to 200 Scam Tracker descriptions (gt_bbb). We use gt_bbb to validate our classifiers.

IV. DATA POLLUTION IN ABUSE REPORTING SERVICES

This section analyzes data pollution in abuse reporting services. We focus on BitcoinAbuse because it did not validate the submitted reports. Additionally, we examine which of the identified issues affect Scam Tracker.

A. Spam Reports

Section III-B identified two prevalent spam report types in BitcoinAbuse: advertisements for funds recovery services (*spam_fundsrecovery*) and fake terrorism reports

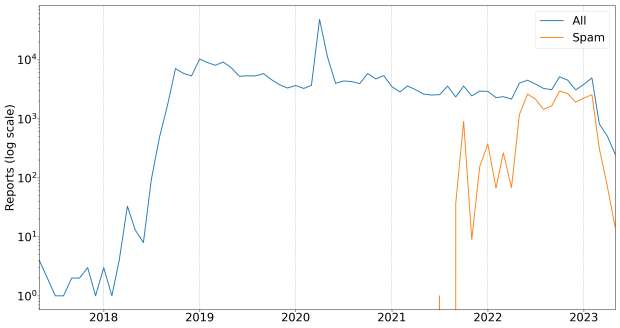


Figure 2. Volume of BitcoinAbuse submitted reports and spam reports (log-scale) per month. Spam reports are those labeled *spam_fundsrecovery* and *spam_terrorism*.

(*spam_terrorism*). In this section, we measure the volume of those spam reports across the whole BitcoinAbuse dataset.

Spammers may post the same spam report hundreds or even thousands of times. For example, the 174 *spam_fundsrecovery* descriptions in gt_final (compared to 16,112 for *sextortion*) appear in 21,070 reports (compared to 17,837 reports for *sextortion*), an average of 121 reports per description (compared to an average of 1.1 for *sextortion*).

To identify *spam_fundsrecovery* reports across all the BitcoinAbuse data, we extract a blocklist of contact indicators from the 174 *spam_fundsrecovery* descriptions in gt_final. The blocklist contains 27 domains, 17 email addresses, 12 phone numbers, 10 URLs, and one Telegram handle. Then, we scan the 287K BitcoinAbuse reports using the blocklist. We conservatively match domains so that if the domain summitrecoup.com is in the blocklist, we match any reports containing the string *summitrecoup*. This process identifies 208 (0.1%) *spam_fundsrecovery* descriptions used in 22,180 (9.9%) reports. To identify *spam_terrorism* reports across all the BitcoinAbuse data, we manually examine the 119 descriptions not in gt_final that contain the string *terror*. This process identifies 79 (0.03%) *spam_terrorism* descriptions used in 1,500 (0.7%) reports.

Figure 2 shows the volume of spam reports (*spam_fundsrecovery*, *spam_terrorism*) and all reports over time. The volume of spam identified is a lower bound, as smaller spam types may also exist. The two prevalent spam types first appeared in February 2021 and dominated all reports since mid-2022, reaching 75% of all submitted reports between June and December 2022. Submitted reports have declined since February 2023 as BitcoinAbuse started redirecting users to ChainAbuse, eventually stopping the

reception of new reports in June 2023.

Our results show that, when no filtering is applied, spam will eventually dominate the submitted reports. In contrast, the Scam Tracker ground truth does not include terrorism spam reports and only a handful of funds recovery reports, indicating that the manual validation filters most spam.

B. Accuracy of User-Reported Abuse Types

Next, we evaluate the accuracy of user-reported BA types by comparing them with the taxonomy abuse types in gt_final. The evaluation is performed on reports because the same description may correspond to multiple reports with different BA types. Thus, we first propagate the taxonomy types from the 19,443 descriptions in gt_final to 43,693 reports (15.2% of all 287,823 reports) where those descriptions appear.

Figure 3 shows a heat map where the heat captures the fraction of the 43K reports assigned the BA type on the y-axis and tagged with the taxonomy abuse type on the x-axis. White means the ratio is zero, light orange means a small ratio, and dark red a high ratio. The bars on top capture the fraction of the 43K reports assigned to each taxonomy abuse type. The bars on the right capture the fraction of the 43K reports assigned to each BA type.

a) Spam impact: Four of the six BA types are dominated by spam reports, namely Bitcoin-Tumbler (98%), Darknet-Market (91%), Ransomware (75%), and Other (43%). For these four BA types, any conclusions drawn without filtering out the spam reports would be heavily biased toward the spam reports and thus would be incorrect.

b) Sextortion: Of the 7,766 reports with the Sextortion BA type, 97.4% are indeed *sextortion*, and 2.5% are *extortion*. Among the 2.5% *extortion* reports, some are likely in reality *sextortion* where the reporter only copied a small part of the received email in the description that did not mention the sexual character of the extortion, making our analysts label it *extortion*. Thus, the Sextortion BA type correctly captures *sextortion* abuse. On the other hand, the most common misclassification (after excluding spam reports) is 22% Ransomware reports that are instead *sextortion*. A possible reason is that reporters who did not know or did not care about the abuse type just selected the first option in the drop-down list (Ransomware). We conclude that reporters could correctly identify *sextortion* abuse. However, most reporters do not care about selecting the right abuse type, which indicates that this task needs to be automated or performed by human analysts.

c) Ransomware: gt_final has no *ransom* reports, but there are 196 *deleteransom* reports. Of those, 173 (88.3%) are correctly assigned the Ransomware BA type. Unfortunately, *deleteransom* is only 1% of Ransomware reports. Thus, the Ransomware BA type is untrustworthy, even after removing spam reports. We conclude that BitcoinAbuse was not commonly used to report ransomware, but when it was (e.g., *deleteransom*), users identified accurately this abuse type.

d) Blackmail-Scam: This BA type was too broad, covering all taxonomy entries underneath *extortion* and *scam*. Of the 6,830 Blackmail-Scam reports, 92% are *sextortion*. Of these, 19% correspond to reports before the introduction of the Sextortion BA type in February 2019, when Blackmail-Scam was the best option for *sextortion*. In the rest, the reporter may have considered that Blackmail-Scam fit the bill and perhaps did not even realize there was another more specific Sextortion entry further down the list. We conclude that broad abuse types should be avoided as they may cause errors due to overlapping more specific abuse types.

e) Bitcoin-Tumbler and Darknet-Market: All reports in these BA types (even after excluding spam reports) were incorrectly classified since gt_final does not have abuse types equivalent to Bitcoin-Tumbler and Darknet-Market. This makes sense as users of mixers and darknet markets may not be inclined to report the services they use.

f) Other: Among the 3,051 Other reports, 43.1% are spam reports, 24.0% are *giveaway*, 23.6% are *investment*, and 9% are scattered across 14 abuse types. The widespread presence of *giveaway* and *investment* scams may indicate that many reporters know how to identify those abuse types. However, since the platform did not offer such types to its users, they opted for the generic type Other. We conclude that it is hard to predict what abuse types victims will report and, thus, which types the online form should offer.

g) Impact on previous works: Previous works analyzing BitcoinAbuse reports rely on the BA types [26], [27], [28]. Unfortunately, we have shown that user-reported BA types are most often incorrect. Conclusions drawn from the Bitcoin-Tumbler, Darknet-Market, and Ransomware BA types will be incorrect as those BA types are dominated by spam, with the remaining reports being mostly misclassifications. On the other hand, conclusions drawn from Sextortion should be trustworthy as most reporters accurately use that BA type for *sextortion* abuse. Blackmail-Scam and Other are too broad. Blackmail-Scam is dominated by *sextortion*, so conclusions drawn on this BA type should match those of Sextortion.

Scam Tracker places all cryptocurrency abuse reports under the same category. Thus, we cannot evaluate the accuracy of the abuse type selection by analysts.

C. Benign Addresses Reported

BitcoinAbuse indexed reports by the reported Bitcoin address. Next, we check whether users may report addresses known to be benign, e.g., belonging to services like exchanges. For this, we leverage four public address-tagging sources [33], [19], [34], [35]. Of the 92,151 reported Bitcoin addresses, 91 (0.1%) are tagged in those sources as internal service addresses or belonging to benign organizations. These are mainly hot wallets used for the daily operations of services and cold wallets used for long-term storage of the service reserves. Of the 91 addresses, 78 belong to exchanges, 5 to mining pools, 3 to the FBI, 2 to payment-processors, 2 to gambling services, and one to a lending provider.

These 91 addresses have 780 reports. Of those, 436 (56%) are *spam_terrorism* claiming that an exchange is involved in terrorism and reporting the hot wallet of the exchange. Another 123 (16%) are *spam_fundsrecovery* posted on frequently reported addresses where they are more likely to be viewed by victims, i.e., the reported address is unrelated to the funds recovery service. The other 28% of reports have two main reasons. First, victims may try to track down where their stolen funds are going and end up reporting an exchange's hot wallet that received the money after the attackers cashed out. Victims should be educated to only report addresses they directly sent funds to (i.e., payment addresses), without trying

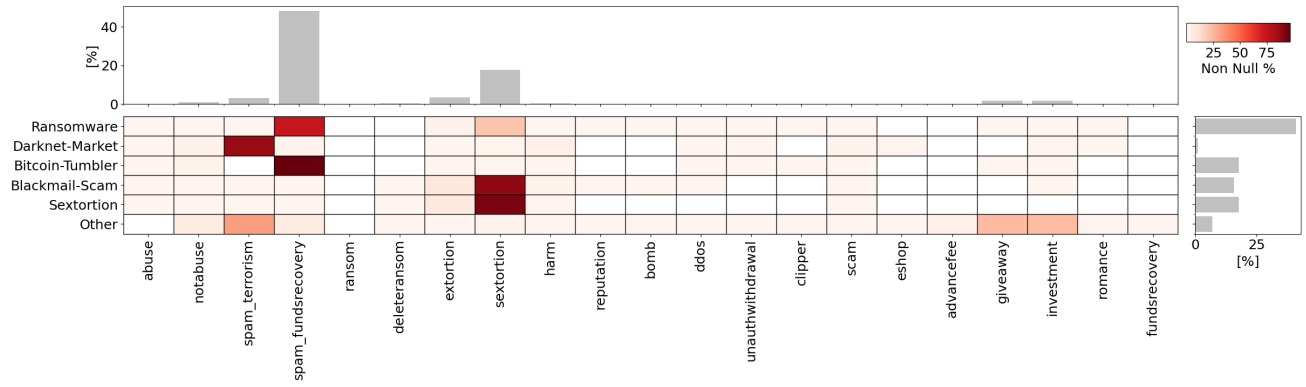


Figure 3. Heat-map capturing the fraction of the 43K evaluation reports assigned the BA type on the y-axis and tagged with the taxonomy abuse type on the x-axis. White means the ratio is zero, light orange is a small ratio, and dark red is a high ratio.

to track down the funds, since information in the blockchain is immutable and tracking can be done later by experienced analysts. In other cases, the reporter may want to pollute an exchange's reputation similar to *spam_terrorism*, e.g., a report asks for a donation to a cold wallet of the HTX exchange.

The Scam Tracker form does not have a field to report blockchain addresses. We extracted 96 Bitcoin addresses from the descriptions and none of those are benign.

a) Impact on revenue estimation: Internal service addresses may receive large amounts of bitcoins unrelated to cybercrime. Thus, any revenue estimation that does not exclude them can hugely overestimate cybercrime revenue. To evaluate this impact, we use WatchYourBack [35] to compute the amount of BTC received by each reported address from the beginning of the Bitcoin blockchain until block 838,800 (April 11, 2024). We calculate a *DD-DC* estimation per address, which filters out self-deposits to avoid double counting [36]. We use the BTC conversion rate to USD on the day of each deposit. Of the 92,151 reported addresses, 58,218 (63.2%) have no transactions until April 11, 2024. Of the rest, 21,429 (23.2%) addresses have received at most $10K, 10,441 (11.3%) between $10K and $1M, 1,565 (1.7%) between $1M and $100M, and 498 (0.5%) over $100M. The 33,933 addresses with transactions have collectively received 123,655,891 BTCs, corresponding to $2,477,761,387,918. When excluding the 91 internal service addresses, the estimation reduces to 48,682,682 BTC ($950,739,558,245). Thus, even if benign addresses are only 0.1% of reported addresses, their impact on cybercrime revenue estimation is overwhelming. Excluding them reduces the estimation by 60% from $2.4 trillion to $950 billion. The latter is still likely an overestimation since additional internal service addresses unknown to our four tagging sources could remain unfiltered.

V. Related Work

a) Cryptocurrency abuse.: Previous works have analyzed a wealth of cryptocurrency abuse types, including ransomware [1], [2], [3], [4], [5], [6], sextortion [7], scams [9], [10], [11], [12], [13], clippers [35], cryptojacking [37], [38], hidden marketplaces [39], [40], [41], human trafficking [42], money laundering [43], and thefts [10]. These works analyze a single abuse type, whereas our work analyzes 19.

b) BitcoinAbuse.: Previous works analyzing BitcoinAbuse report on the presence of heavy reporters [26], the concentration of revenue on a few addresses [27], and the revenue per abuse type [28]. However, they did not consider data pollution: they did not filter out spam reports, considered all reported addresses to be malicious (and thus their deposits to be cybercriminal revenue), and assumed reporter-selected abuse types were correct. Our results cast a shadow over their conclusions.

VI. Conclusion

We have analyzed 289K abuse reports from two services to measure the extent and impact of data pollution. For this, we have built a public ground truth dataset of 19,443 abuse descriptions labeled with 19 popular abuse types and have used it to show that, if left unchecked, spam reports will eventually flood an abuse reporting service; that a small fraction of reported addresses may be benign (0.1%) but those are responsible for a majority (60%) of the deposited funds; and that user-reported abuse types cannot be trusted. We are currently using the produced datasets to design automated approaches for the identification of useful reports and their classification into abuse types.

Funding

This work was partially funded by the Spanish Government MCIN/AEI/10.13039/501100011033/ through grants TED2021-132464B-I00 (PRODIGY), PID2022-142290OB-I00 (ESPADA), PRE2019-088472, and PREP2022-000165. The work of NEC was partially funded by the European Union's Horizon Europe program under Grant Agreement No. 101139285 (NATWORK) and by the Smart Networks and Services Joint Undertaking (SNS JU). The above grants are co-funded by European Union ESF, EIE, and NextGeneration funds.

References

[1] M. Spagnuolo, F. Maggi, and S. Zanero, "BitIodine: Extracting Intelligence from the Bitcoin Network," in *Financial Cryptography and Data Security*, 2014.

[2] K. Liao, Z. Zhao, A. Doupé, and G.-J. Ahn, "Behind Closed Doors: Measurement and Analysis of CryptoLocker Ransoms in Bitcoin," in *APWG Symposium on Electronic Crime Research*, June 2016.

[3] M. Conti, A. Gangwal, and S. Ruj, "On the economic significance of ransomware campaigns: A bitcoin transactions perspective," *Computers & Security*, vol. 79, pp. 162–189, 2018.

[4] D. Y. Huang, M. M. Aliapoulios, V. G. Li, L. Invernizzi, K. McRoberts, E. Bursztein, J. Levin, K. Levchenko, A. C. Snoeren, and D. McCoy, "Tracking Ransomware End-to-end," in *IEEE Symposium on Security and Privacy*, May 2018.

[5] M. Paquet-Clouston, B. Haslhofer, and B. Dupont, "Ransomware Payments in the Bitcoin Ecosystem," *Journal of Cybersecurity*, vol. 5, no. 1, 2019.

[6] K. van Liebergen, G. Gomez, S. Matic, and J. Caballero, "All your (data)base are belong to us: Characterizing Database Ransom(ware) Attacks," in *Network and Distributed Systems Security Symposium*, 2025.

[7] M. Paquet-Clouston, M. Romiti, B. Haslhofer, and T. Charvat, "Spams Meet Cryptocurrencies: Sextortion in the Bitcoin Ecosystem," in *ACM Conference on Advances in Financial Technologies*, 2019.

[8] F. Oggier, A. Datta, and S. Phetsouvanh, "An ego network analysis of sextortionists," *Social Network Analysis and Mining*, vol. 10, pp. 1–14, 2020.

[9] M. Bartoletti, B. Pes, and S. Serusi, "Data Mining for Detecting Bitcoin Ponzi Schemes," in *Crypto Valley Conference on Blockchain Technology*, June 2018.

[10] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, "A Fistful of Bitcoins: Characterizing Payments among Men with No Names," in *Internet Measurement Conference*, 2013.

[11] P. Xia, H. Wang, B. Zhang, R. Ji, B. Gao, L. Wu, X. Luo, and G. Xu, "Characterizing Cryptocurrency Exchange Scams," *Computers & Security*, vol. 98, 2020.

[12] M. Bartoletti, S. Lande, A. Loddo, L. Pompianu, and S. Serusi, "Cryptocurrency Scams: Analysis and Perspectives," *IEEE Access*, vol. 9, pp. 148 353–148 373, 2021.

[13] X. Li, A. Yepuri, and N. Nikiforakis, "Double and nothing: Understanding and detecting cryptocurrency giveaway scams," in *Network and Distributed Systems Security Symposium*, 2023.

[14] "Bitcoin abuse," 2023, https://www.bitcoinabuse.com.

[15] "Chainabuse," 2024, https://chainabuse.com.

[16] "Scamtracker," 2024, https://www.bbb.org/scamtracker.

[17] "Scamwatch," 2024, https://www.scamwatch.gov.au.

[18] "Reportfraud," 2024, https://reportfraud.ftc.gov.

[19] "GlassChain," 2024, https://glasschain.org.

[20] "Crystal Lite," 2024, https://lite.crystalintelligence.com.

[21] F. T. Commission, "Consumer sentinel network - data book 2023," February 2024, https://www.ftc.gov/system/files/ftc_gov/pdf/CSN-Annual-Data-Book-2023.pdf.

[22] Better Business Bureau Institute for Marketplace Trust, "BBB Scam Tracker risk report," April 2024, https://bbbmarketplacetrust.org/wp-content/uploads/2024/04/2023-BBBScamTracker-RiskReport-US-040224.pdf.

[23] M. Bitaab, H. Cho, A. Oest, Z. Lyu, W. Wang, J. Abraham, R. Wang, T. Bao, Y. Shoshitaishvili, and A. Doupé, "Beyond Phish: Toward Detecting Fraudulent e-Commerce Websites at Scale," in *IEEE Symposium on Security and Privacy*, 2023.

[24] P. Kotzias, K. Roundy, M. Pachilakis, I. Sanchez-Rola, and L. Bilge, "Scamdog Millionaire: Detecting E-commerce Scams in the Wild," in *Annual Computer Security Applications Conference*, 2023.

[25] P. Kotzias, M. Pachilakis, J. Aldana Iuit, J. Caballero, I. Sanchez-Rola, and L. Bilge, "Ctrl+Alt+Deceive: Quantifying User Exposure to Online Scams," in *Network and Distributed Systems Security Symposium*, 2025.

[26] J. Choi, J. Kim, M. Song, H. Kim, N. Park, M. Seo, Y. Jin, and S. Shin, "A large-scale bitcoin abuse measurement and clustering analysis utilizing public reports," *IEICE TRANSACTIONS on Information and Systems*, vol. 105, no. 7, pp. 1296–1307, 2022.

[27] D. Buil-Gil and P. Saldaña-Taboada, "Offending concentration on the internet: An exploratory analysis of bitcoin-related cybercrime," *Deviant Behavior*, vol. 43, no. 12, pp. 1453–1470, 2022.

[28] H. Rosenquist, D. Hasselquist, M. Arlitt, and N. Carlsson, "On the Dark Side of the Coin: Characterizing Bitcoin Use for Illicit Activities," in *International Conference on Passive and Active Network Measurement*, 2024.

[29] "langdetect," 2023, https://pypi.org/project/langdetect.

[30] "iocsearcher," 2023, https://github.com/malicialab/iocsearcher.

[31] M. Vasek and T. Moore, "There's no free lunch, even using bitcoin: Tracking the popularity and profits of virtual currency scams," in *Financial Cryptography and Data Security*, 2015.

[32] E. Badawi and G.-V. Jourdan, "Cryptocurrencies emerging threats and defensive mechanisms: A systematic literature review," *IEEE Access*, vol. 8, pp. 200 021–200 037, 2020.

[33] "Arkham Intel," 2024, https://intel.arkm.com.

[34] B. Haslhofer, R. Karl, and E. Filtz, "O Bitcoin Where Art Thou? Insight into Large-Scale Transaction Graphs," in *SEMANTiCS Conference*, 2016.

[35] G. Gomez, P. Moreno-Sanchez, and J. Caballero, "Watch your back: Identifying cybercrime financial relationships in bitcoin through back-and-forth exploration," in *ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[36] G. Gomez, K. van Liebergen, and J. Caballero, "Cybercrime bitcoin revenue estimations: Quantifying the impact of methodology and coverage," in *ACM SIGSAC Conference on Computer and Communications Security*, 2023.

[37] D. Y. Huang, H. Dharmdasani, S. Meiklejohn, V. Dave, C. Grier, D. McCoy, S. Savage, N. Weaver, A. C. Snoeren, and K. Levchenko, "Botcoin: Monetizing Stolen Cycles," in *Network and Distributed Systems Security Symposium*, 2014.

[38] E. Tekiner, A. Acar, A. S. Uluagac, E. Kirda, and A. A. Selcuk, "SoK: Cryptojacking Malware," in *IEEE European Symposium on Security and Privacy*, 2021.

[39] N. Christin, "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace," in *The World Wide Web Conference*, 2013.

[40] S. Lee, C. Yoon, H. Kang, Y. Kim, Y. Kim, D. Han, S. Son, and S. Shin, "Cybercriminal Minds: An Investigative Study of Cryptocurrency Abuses in the Dark Web," in *Network and Distributed Systems Security Symposium*, 2019.

[41] D. Ron and A. Shamir, "How Did Dread Pirate Roberts Acquire and Protect his Bitcoin Wealth?" in *Financial Cryptography and Data Security*, 2014.

[42] R. S. Portnoff, D. Y. Huang, P. Doerfler, S. Afroz, and D. McCoy, "Backpage and Bitcoin: Uncovering Human Traffickers," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[43] M. Möser, R. Böhme, and D. Breuker, "An Inquiry into Money Laundering Tools in the Bitcoin Ecosystem," in *APWG eCrime Researchers Summit*, September 2013.

Una propuesta para mejorar la seguridad de las hardware wallets

Miquel Calonge Fabregat
Dto. Ciencias de la Computación
Universidad de Alcalá
Av. Principal de la Universidad, 1E
28805 Alcalá de Henares
miquel.calonge@uah.es

Jose Amelio Medina Merodio
Dto. Ciencias de la Computación
Universidad de Alcalá
Av. Principal de la Universidad, 1E
28805 Alcalá de Henares
josea.medina@uah.es

Julio Hernández Castro
Dto. Sistemas Informáticos
Universidad Politécnica de Madrid
C./ Alan Turing, s/n
28031 Madrid
jc.hernandez.castro@upm.es

***Resumen*—Este trabajo analiza el estado del arte en la seguridad de las hardware wallets, dispositivos utilizados para custodiar y gestionar criptoactivos. Se revisan algunas investigaciones previas, con el objetivo primario de identificar áreas en las que la seguridad pueda ser mejorada sin comprometer las restricciones de hardware que generalmente presentan estos dispositivos. Uno de los aspectos detectados como crítico es la total falta de mecanismos para verificar si la wallet sincronizada es vulnerable. Actualmente, los dispositivos estudiados no realizan ninguna comprobación en este sentido. En respuesta, el artículo propone una solución de muy fácil implementación que mejora sensiblemente la seguridad de las hardware wallets sin afectar significativamente su rendimiento o usabilidad. En concreto, proponemos la incorporación de un sistema de verificación automatizado que evalúe si la wallet cumple con ciertos requisitos mínimos de seguridad antes de proceder a su sincronización.**

Tipo de contribución: *Investigación original*

I. Introducción

Existen diversos métodos para la gestión de criptoactivos, cada uno con características específicas en términos de seguridad, coste y usabilidad. Uno de los más básicos es el uso de paper wallets, que consisten en la impresión física de las claves privadas en papel. Aunque esta opción ofrece un alto grado de protección frente a ataques digitales, presenta riesgos significativos en cuanto a pérdida, deterioro o robo del soporte físico, lo que puede comprometer el acceso a los fondos custodiados.

Otra alternativa ampliamente utilizada es el empleo de billeteras basadas en software, como las extensiones de navegador, entre las que destaca MetaMask. Estas soluciones permiten a los usuarios gestionar sus activos digitales de manera sencilla y acceder a aplicaciones descentralizadas. Sin embargo, al estar constantemente conectadas a Internet, son más vulnerables a ataques de phishing, malware o fallos de seguridad en el navegador y/o sistema operativo.

Asimismo, las billeteras móviles son un ejemplo de billeteras software que han ganado popularidad debido a su facilidad de uso y portabilidad en dispositivos personales. Aplicaciones de este tipo permiten gestionar criptoactivos de manera rápida y cómoda, pero también presentan riesgos asociados a la exposición a malware, la pérdida o robo del dispositivo y la posibilidad de ataques de intermediario (MITM).

Finalmente, las hardware wallets, objeto principal de estudio de este artículo, son también un método popular para la gestión y custodia de criptoactivos principalmente por su usabilidad y la confianza que genera disponer de un dispositivo físico. Dado que una gran cantidad de personas e instituciones emplea estos dispositivos para administrar sus fondos, el volumen de capital gestionado a través de ellos es muy elevado. Por ello, su seguridad es un aspecto crítico que requiere especial atención, ya que cualquier vulnerabilidad en su diseño o implementación podría comprometer la integridad de activos significativos.

II. Estado del arte

A lo largo de esta sección, se analizan las principales investigaciones recientes sobre la seguridad en hardware wallets. Se han agrupado los trabajos en cuatro áreas principales: vulnerabilidades y ataques, protección de claves y autenticación, integración con blockchain, y amenazas emergentes. Esta revisión proporciona un panorama de los desafíos actuales y las innovaciones que buscan fortalecer la seguridad de estos dispositivos esenciales para la protección de activos digitales.

II-A. Vulnerabilidades y ataques a hardware wallets

Las hardware wallets, aunque diseñadas para ser altamente seguras, no están exentas de vulnerabilidades. En los últimos años, se han identificado múltiples ataques que explotan debilidades tanto en su hardware como en su software o firmware. Los ataques pueden clasificarse en tres categorías principales: ataques físicos, ataques de canal lateral y ataques de software/malware.

II-A1. Ataques físicos a hardware wallets: Uno de los ataques más relevantes fue el identificado en [1], quienes analizaron la seguridad de la EEPROM DS28C36, utilizada en hardware wallets como la Coldcard Mk4. Su investigación demostró que, aunque el chip resiste ataques de falla única, es vulnerable a inyecciones de fallos con doble láser. Este tipo de ataque permite extraer información confidencial mediante la alteración de los circuitos internos del dispositivo. Este hallazgo subraya la importancia de implementar protecciones contra fallas inducidas en hardware wallets.

Otros ataques físicos incluyen la manipulación del hardware antes de la entrega al usuario (supply chain attacks), en los cuales un atacante introduce hardware malicioso en la wallet antes de que el usuario la reciba. Esto se ha convertido en un problema creciente, dado que muchos dispositivos se venden a través de terceros y pueden ser interceptados y manipulados con relativa facilidad.

ISBN:978-84-10169-61-6

II-A2. Ataques de canal lateral: Los ataques de canal lateral explotan información derivada de la ejecución de operaciones criptográficas, como el consumo de energía, el tiempo de ejecución y la radiación electromagnética emitida. [2] demuestra que es posible extraer claves privadas observando el consumo de energía durante operaciones criptográficas.

[3] va más allá al demostrar un ataque basado en la observación del consumo energético en la multiplicación escalar en curvas elípticas, una operación clave en la generación y uso de claves privadas. Su método, validado en hardware wallets populares como Trezor Model One, demostró que con suficiente análisis es posible recuperar la clave privada con un solo trazo energético, lo que representa una amenaza muy significativa para la seguridad de estos dispositivos.

II-A3. Ataques de software y malware: Los ataques basados en software representan una amenaza creciente, ya que permiten comprometer la operación de hardware wallets sin necesidad de acceso físico al dispositivo. Un ejemplo de ello es el malware EthClipper, identificado en [4], que modifica direcciones de criptomonedas copiadas al portapapeles, engañando a los usuarios para que envíen fondos a direcciones controladas por atacantes. Este malware logra evadir los mecanismos de verificación visual de hardware wallets como Ledger, Trezor y KeepKey con un 50 % de éxito en ataques que simulan direcciones similares a las reales.

Otro ataque relevante es el desarrollado en [5], quienes crearon herramientas forenses que exploran la memoria de hardware wallets como Ledger Nano S y Electrum. Su investigación demostró que es posible extraer claves privadas cifradas y utilizar ataques de diccionario masivos para descifrarlas.

II-B. Protección de claves, autenticación y seguridad en blockchain

Dado el creciente número de ataques dirigidos a hardware wallets, la investigación en nuevos mecanismos de protección de claves y autenticación segura ha cobrado una importancia crucial [6].

II-B1. Uso de hardware seguro para proteger claves privadas: Una de las soluciones más prometedoras ha sido la integración de módulos de seguridad de hardware (HSM) en hardware wallets. [7-8] propone combinar HSM con tecnologías blockchain en entornos IoT para mejorar la seguridad de las transacciones. En su modelo, cada nodo de la red utiliza HSM para almacenar claves privadas de manera segura y garantizar que las transacciones sean firmadas únicamente por dispositivos confiables.

Otra propuesta innovadora es CryptoVault, desarrollado en [9], que emplea Intel SGX para mantener las claves privadas en un enclave seguro. Este método evita la exposición de claves incluso si el sistema es comprometido.

II-B2. Métodos avanzados de autenticación y respaldo de claves: Para abordar el problema de la pérdida de claves privadas, [10] desarrolla un sistema de respaldo basado en verificación visual humana, que permite recuperar claves sin depender de frases mnemotécnicas, reduciendo el riesgo de exposición a ataques de phishing o malware.

De manera similar, [11] explora la posibilidad de integrar hardware wallets con entornos de ejecución de confianza (TEE) en teléfonos celulares, permitiendo una gestión más segura de las claves sin comprometer la facilidad de uso.

II-C. Integración de Blockchain y aplicaciones de seguridad en hardware wallets

Las hardware wallets también han sido exploradas en el contexto de aplicaciones más amplias dentro de la seguridad en blockchain. [12] propone un mecanismo de protección de datos en endpoints de blockchain mediante hardware wallets que implementan privacidad transaccional con hashing de Blake3.

Además, [13] desarrolla HistoTrust, un esquema que combina hardware seguro y blockchain para rastrear el comportamiento de la inteligencia artificial, proporcionando confianza en la trazabilidad de la información. De manera similar, [14] explora el almacenamiento distribuido dinámico para la escalabilidad de blockchain, mejorando la integridad de los datos sin comprometer su seguridad.

En el ámbito de IoT, [15] diseña una billetera de criptomonedas segura basada en una versión mejorada del algoritmo ECDSA. Su solución combina hardware wallets, blockchain y protocolos de seguridad mejorados para garantizar transacciones seguras en el entorno IoT.

Estos estudios reflejan cómo las hardware wallets están evolucionando más allá del almacenamiento seguro de claves privadas, incorporando blockchain y tecnologías avanzadas para garantizar una mayor seguridad en aplicaciones distribuidas.

II-D. Amenazas emergentes y nuevas soluciones

A medida que los ataques a hardware wallets evolucionan, también surgen nuevas estrategias para fortalecer su seguridad. [16] evalúa la viabilidad de ataques de fuerza bruta sobre claves públicas de criptomonedas, concluyendo que, aunque poco rentables, aún representan una amenaza potencial que debe ser estudiada.

En el ámbito de la autenticación, [17] propone un mecanismo basado en criptografía postcuántica y funciones no clonables (PUF) para autenticar drones en redes IoT, una propuesta que con algunos cambios podría aplicarse a hardware wallets en el futuro. [18] realiza un análisis de gestión de claves en custodios de criptomonedas, discutiendo la efectividad de las hardware wallets en escenarios de custodia segura.

Por otro lado, [19] realiza el primer análisis forense detallado de la memoria de varias hardware wallets, revelando que se pueden extraer claves públicas y otros artefactos que podrían desanonimizar transacciones pasadas y futuras.

Estos estudios enfatizan la necesidad de una vigilancia constante y de la adopción de nuevas tecnologías para mitigar las amenazas emergentes en el uso de hardware wallets.

III. Características y opciones de uso de una hardware wallet

En esta sección se explora la usabilidad de las hardware wallets, centrándonos en las opciones principales que suelen ofrecer al usuario. Se analizarán los procesos de generación de una nueva billetera, la restauración de una billetera existente y el uso de una passphrase como una capa adicional de seguridad.



Figura 1. Algunos de los hardware wallets más utilizados. A la izquierda, tres modelos de Trezor. A la derecha, una Ledger Nano S. Mas informacion en https://trezor.io y https://www.ledger.com/

El uso de una hardware wallet implica una serie de interacciones específicas que la diferencian de otros métodos de almacenamiento de claves privadas. Una hardware wallet está diseñada para minimizar la exposición de la clave privada a posibles amenazas. Por este motivo, la interfaz de usuario suele ser minimalista, generalmente reducida a una pequeña pantalla y pocos botones físicos que permiten realizar acciones fundamentales como confirmar transacciones o navegar por el menú del dispositivo.

Otro aspecto relevante es que, aunque estos dispositivos tienen una curva de aprendizaje inicial, su uso se ha simplificado con los años. La mayoría de hardware wallets actuales cuentan con una configuración guiada y utilizan estándares como BIP-39 para la recuperación de claves, facilitando su uso tanto para usuarios experimentados como para aquellos que están ingresando al ecosistema de criptomonedas por primera vez.

A continuación, se detallarán las principales interacciones que los usuarios pueden realizar con una hardware wallet y las implicaciones de seguridad de cada una.

III-A. Generación de una nueva wallet

Una de las primeras acciones que un usuario puede realizar con una hardware wallet es la generación de una nueva billetera. Para ello, el dispositivo incorpora un Pseudo-Random Number Generator (PRNG), encargado de generar una semilla única que servirá como base para la billetera. Esta semilla se transforma en una serie de palabras mnemotécnicas (generalmente 12, 18 o 24 palabras) siguiendo el estándar BIP-39, facilitando su almacenamiento seguro y recuperación futura.

Es importante destacar que la generación de aleatoriedad en hardware wallets es un desafío técnico en sí mismo, ya que estos dispositivos suelen contar con recursos computacionales bastante limitados en comparación con los de un ordenador convencional. Para mejorar la calidad de la aleatoriedad, muchas hardware wallets combinan múltiples fuentes de entropía.

Además de generar la semilla, el usuario deberá anotar y guardar estas palabras mnemotécnicas en un lugar seguro, ya que son la única forma de recuperar la billetera en caso de pérdida o daño del dispositivo. Algunas hardware wallets ofrecen opciones adicionales, como el uso de una tarjeta de respaldo metálica para proteger la semilla contra el deterioro físico.

El código de las hardware wallets suele ser open source, por lo que se encuentra disponible para revisión pública. Tras analizar algunas implementaciones, se ha observado que en ningún caso existe una capa adicional de seguridad que verifique si la wallet generada por el dispositivo es aparentemente fuerte o si ha ocurrido algún error durante el proceso de inicialización o generación. Estos errores son comunes en caso de problemas en el proceso de manufactura, distribución o inicialización. Controlar este aspecto representa una potencial mejora que, sorprendentemente, no se ha tratado en investigaciones previas.

III-B. Restauración o selección manual de una wallet

Otra funcionalidad clave de una hardware wallet es la capacidad de restaurar una billetera existente. Esto permite a los usuarios recuperar sus fondos en un nuevo dispositivo en caso de que el original se pierda o deje de funcionar. Este proceso consiste en ingresar manualmente la semilla previamente generada y, a partir de ella, que la hardware wallet regenere las claves y direcciones asociadas.

Aunque la función de restauración se presenta con este propósito, en la práctica también permite a los usuarios ingresar cualquier semilla válida y, de este modo, elegir manualmente una billetera específica. Esto puede ser útil en escenarios donde una persona desee mover la gestión a una wallet generada previamente en otro entorno, o pretenda, aunque no es aconsejable, elegir los componentes de su propio conjunto de palabras mnemotécnicas.

La restauración manual también implica riesgos, especialmente si la semilla ha sido almacenada en un medio digital, comprometida de alguna manera o generada con algún tipo de debilidad.

Adicionalmente, algunas hardware wallets modernas incluyen soporte para múltiples cuentas y direcciones en un solo dispositivo, permitiendo a los usuarios administrar diversas billeteras sin necesidad de restaurar manualmente una semilla diferente cada vez.

Como para la funcionalidad del apartado anterior, para esta también se han analizado algunas implementaciones y se ha observado exactamente la misma problemática. En ningún caso se han observado medidas que comprueben si la seed restaurada por el usuario pueda tener algún tipo de debilidad. Aquí, la problemática es mucho mayor, ya que es bastante más probable que se configure una billetera débil de forma manual, que la genere el propio dispositivo.

III-C. Passphrase: una capa extra de seguridad

Una característica de algunas hardware wallets es el uso de una passphrase. A diferencia de la generación de la semilla, que es gestionada por el PRNG del dispositivo, la passphrase es una entrada manual del usuario que altera la derivación de las claves, generando una billetera completamente distinta.

El uso de passphrase proporciona una capa adicional de seguridad, ya que un atacante que logre obtener la semilla de una wallet seguirá sin acceso a los fondos si no conoce la passphrase correspondiente. Aunque la clave privada sigue perteneciendo al mismo espacio de búsqueda, en la práctica, la seguridad aumenta porque el atacante necesita conocer tanto la semilla como la passphrase correcta para acceder a los fondos.

Existen distintas estrategias de uso de passphrase. Algunos usuarios configuran una billetera principal sin passphrase y otra con una passphrase específica, permitiendo el uso de la estrategia de ”plausible deniability”(negación plausible). En caso de coacción, el usuario puede revelar la billetera sin passphrase mientras mantiene sus verdaderos fondos protegidos detrás de una passphrase desconocida para el atacante.

Algorithm 1 Como derivar una nueva dirección Bitcoin a partir de una clave privada + passphrase

procedure DERIVBTCADDRESS($priv_key$, $passphrase$)
  // Combinar priv+passphrase con HMAC-SHA256
  $combined \leftarrow$ HMAC-SHA256$(priv_key, passphrase)$
  // Derivar nueva privada (ejemplo: XOR)
  $new_priv_key \leftarrow priv_key \oplus combined$
  // Generar pública (secp256k1) con nueva privada
  $pub_key \leftarrow$ EllipticCurveMult(new_priv_key, G)
  ▷ G un generador
  // Doble hash a la clave pública:
  $hash_pub \leftarrow$ RIPEMD160(SHA256(pub_key))
  // Agregar prefijo de red (0x00 para BTC mainnet)
  $ph \leftarrow$ Concatenate$(0x00, hash_pub)$
  // Calcular checksum:
  $checksum \leftarrow$ First4Bytes(SHA256(SHA256(ph)))
  // Concatenar el hash prefijado con el checksum
  $binary_address \leftarrow$ Concatenate$(ph, checksum)$
  // Codificar el resultado en Base58
  $btc_address \leftarrow$ Base58Encode$(binary_address)$
  return $btc_address$
end procedure

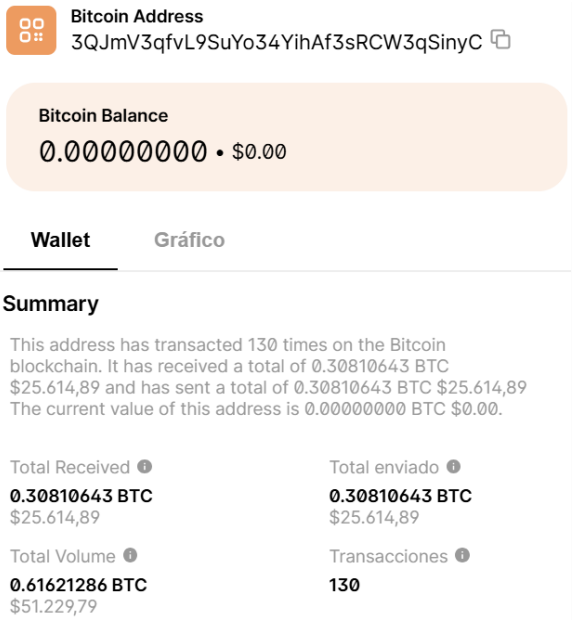


Figura 2. Ejemplo de la actividad de una billetera vulnerable.

Sin embargo, el uso de passphrase también tiene sus riesgos. A diferencia de la semilla, que suele anotarse en papel o en medios físicos resistentes, la passphrase no se muestra en el dispositivo ni se almacena en la hardware wallet. Esto significa que si el usuario olvida su passphrase, perderá acceso definitivo a los fondos asociados a esa billetera. Por ello, se recomienda utilizar una passphrase que sea segura pero recordable o emplear mecanismos de almacenamiento seguros para respaldarla sin comprometer su seguridad.

En resumen, el uso de passphrase es una medida opcional, pero altamente recomendada para mejorar la seguridad de las hardware wallets. Bien utilizada, permite a los usuarios reforzar la protección de sus fondos y gestionar múltiples billeteras en un mismo dispositivo de manera más segura.

IV. ANÁLISIS DE CRITICIDAD

Como se ha mencionado previamente, existe una importante deficiencia de seguridad en muchas hardware wallets en lo referente a la ausencia total de tests posteriores a la creación o restauración de una billetera. Este aspecto puede sorprender, especialmente dada la criticidad de estas herramientas, específicamente diseñadas para custodiar enormes balances de fondos digitales, con frecuencia a nivel institucional.

Esta situación puede analizarse en comparación con los mecanismos típicos empleados por la mayoría de aplicaciones y páginas web durante el registro o creación de una cuenta de usuario. En estos casos, es habitual que las contraseñas ingresadas por los usuarios se contrasten con bases de datos internas o externas de contraseñas débiles (típicamente mediante librerías) o comprometidas (mediante consultas en bases de datos). Herramientas como la librería check-password-strength[1] o la web haveibeenpwned.com [2] son ejemplos de estos servicios y herramientas. Afortunadamente, este tipo de verificaciones son ahora una práctica estándar para reforzar la seguridad de los usuarios, rechazando inmediatamente contraseñas que se consideran inseguras o de poca complejidad. Las webs modernas suelen exigir combinaciones complejas que incluyen letras mayúsculas, minúsculas, números y caracteres especiales, precisamente porque estas prácticas disminuyen significativamente el riesgo de que las cuentas sean vulneradas por ataques de fuerza bruta o ataques basados en diccionarios.

La ausencia de cualquier comprobación de seguridad en la mayoría de las hardware wallets es realmente preocupante: estas se usan incluso en una plataforma web común, cuando el impacto económico de un ataque a la cuenta de un usuario puede ser mínimo. Incluso en casos en los que hay poco en juego se exigen comprobaciones mínimas de seguridad para aceptar una contraseña, ¿cómo es entonces posible que las hardware wallets, dispositivos creados específicamente para gestionar fondos, permitan la generación o restauración de claves privadas sin una comprobación mínima de seguridad? Esta ausencia es aún más alarmante considerando que algunas claves privadas débiles pueden ser detectadas inmediatamente por una breve implementación de seguridad en un smart contract y de ser utilizadas para recibir fondos, un atacante acabará por sustraerlos, pudiendo producirse ambas transacciones (recepción de los fondos y sustracción de los mismos) habitualmente en el mismo bloque.

En el contexto de una plataforma web tradicional, un atacante no puede directamente vulnerar las cuentas de usuario sin antes haber conseguido acceso a recursos críticos, como bases de datos que contienen hashes o información sensible. Esto implica pasos previos significativos, como obtener acceso inicial al sistema, escalar privilegios o comprometer recursos internos, lo que hace que el ataque dirigido a un sitio específico sea considerablemente difícil y, en muchos casos, inviable. Además, las plataformas pueden implementar múltiples capas de seguridad, lo que aumenta el esfuerzo requerido por parte de los atacantes.

[1] https://github.com/deanilvincent/check-password-strength
[2] https://haveibeenpwned.com/Passwords

En cambio, en el entorno de blockchain, la situación es radicalmente diferente. Numerosos estudios han documentado cómo los atacantes emplean escáneres automatizados que monitorean constantemente la red en busca de claves privadas vulnerables o débiles. Debido a la naturaleza pública y transparente de la blockchain, cualquier dirección con debilidades queda inmediatamente expuesta, y los fondos asociados pueden ser sustraídos instantáneamente. A diferencia del contexto web tradicional, aquí el atacante no necesita obtener previamente acceso a recursos críticos ni superar barreras internas; desde cualquier ubicación y sin obstáculos previos, es posible lanzar un ataque contra direcciones vulnerables. Este análisis y propuesta preliminar deberían servir para elevar la visibilidad de este problema, sensibilizando a los creadores de hardware wallets para que introduzcan medidas de seguridad extra.

V. Propuesta de mejora

Antes de abordar las posibles soluciones a esta problemática, hay que considerar que una comprobación extensiva puede presentar ciertas limitaciones inherentes a la naturaleza del hardware wallet. Una hardware wallet está diseñada explícitamente para operar aislada del entorno externo, minimizando al máximo posible la superficie de ataque. Debido a esta característica, estos dispositivos no deben depender de recursos externos como bases de datos en línea o consultas remotas para verificar la seguridad o robustez de claves privadas generadas o restauradas. La razón principal de esta restricción es la seguridad misma; conectar la hardware wallet a bases de datos externas podría introducir vectores de ataque adicionales. Además, mantener una base de datos interna con listas de claves privadas conocidas o potencialmente inseguras sería altamente impráctico debido a limitaciones técnicas, especialmente considerando que estos dispositivos están pensados para tener un almacenamiento mínimo, optimizando su rendimiento y eficiencia energética.

Ante este contexto, una solución viable sería integrar mecanismos de análisis basados en aleatoriedad directamente en el firmware del dispositivo. La aleatoriedad es un concepto clave en criptografía y seguridad informática, relacionada en este caso con el grado de impredecibilidad de una clave. Por ejemplo, un análisis simple podría transformar la clave privada generada a formato binario y evaluar su proporción de ceros y unos. Una alerta podría activarse automáticamente si dicha proporción supera ciertos límites. Este tipo de análisis es sencillo, ligero en términos de recursos, y proporcionaría al usuario una advertencia temprana sobre potenciales vulnerabilidades.

El test que vamos a proponer es muy eficiente y rápido, e intenta medir propiedades de la secuencia binaria a analizar que tienen en cuenta tanto su uniformidad como su independencia. En concreto, combinaremos un estudio de la entropía de la secuencia junto a su coeficiente de correlación.

V-1. Entropia: Este test calcula la entropía tal y como se define en Teoría de la Información. Si llamamos a la secuencia a estudiar $X = \{x_k\}_{k=0}^{N-1}$, el número de ocurrencias i de los valores entre 0 y 15 se cuentan. Llamamos $P_X(i) = \frac{Ap_i}{n}$, donde $Ap_i = \#\{0 \leq k \leq n-1 : x_k = i\}$. La entropía se calcula mediante la fórmula:

$$H(X) = -\sum_{i=0}^{i=15} P_X(i) \log_2 P_X(i). \tag{1}$$

Intuitivamente, la entropía indica la cantidad de bits por nibble de información incomprensible que contiene la secuencia. Por lo tanto, su valor oscila entre 0 y 4. Un valor de 0 significa que la secuencia no contiene información, y 4 significa que es incomprensible.

V-2. Correlacion Serie: Esta prueba mide la dependencia de un nibble con respecto al nibble inmediatamente anterior. Cuenta las apariciones de cada par de nibbles y devuelve un valor entre -1 y 1, donde 0 significa que no hay relación lineal entre nibbles y 1 representa una dependencia lineal completa. Este se calcula como el coeficiente de correlación de Pearson r entre la secuencia $x = x_0...x_{n-1}$ y la secuencia desplazada un nibble "a la izquierda,"$x' = x_1...x_{n-1}x_0$. El coeficiente de correlación de Pearson r entre dos muestras x e y se expresa mediante la siguiente expresión.

$$r = \frac{\sum_{i=0}^{n-1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n-1}(x_i - \bar{x})^2}\sqrt{\sum_{i=0}^{n-1}(y_i - \bar{y})^2}}. \tag{2}$$

que en nuestro caso es

$$r = \frac{\sum_{i=0}^{n-1}(x_i - \bar{x})(x'_i - \bar{x})}{\sum_{i=0}^{n-1}(x_i - \bar{x})^2} = \frac{\sum_{i=0}^{n-1} x_i x'_i - n\bar{x}^2 - n\bar{x}^2 + n\bar{x}^2}{\sum_{i=0}^{n-1}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)} = \frac{\sum_{i=0}^{n-1} x_i x_{(i+1) \bmod n} - n\bar{x}^2}{\sum_{i=0}^{n-1} x_i^2 - n\bar{x}^2}.$$

Algorithm 2 Verificación de adecuacion de Clave Privada

```
1:  procedure VERIFICARCLAVEPRIVADA(clave_privada)
2:      binario ← CONVERTIRABINARIO(clave_privada)
3:      total ← LONGITUD(binario)
4:      for i in (0...15)
5:      contador_i ← CONTAR(binario, 'i')
6:      Run the Entropy, Serial Correlation tests
7:      if Entropy < 3,5 or ||Scc|| > 0,1 then
8:          lanzar Excepción Clave inválida
9:      else
10:         return binario
11:     end if
12: end procedure
```

Asimismo, podrían implementarse otros mecanismos específicos para mitigar debilidades reconocidas relacionadas con estándares como BIP-39, utilizado para generar semillas mnemónicas en hardware wallets. Un ejemplo claro de una buena práctica sería comprobar una mínima diversidad dentro de las palabras generadas. Por ejemplo, que en una semilla estándar de 12 palabras al menos 8 sean únicas entre sí, evitando así semillas con muchas repeticiones, algo que claramente

reduciría su resistencia ante ataques por diccionario. Este test sencillo, como cualquier test de aleatoriedad, sería necesario pero no suficiente para probar la robustez de una clave.

Estos métodos no sólo aumentarían considerablemente la seguridad, sino que además respetarían las limitaciones propias de las hardware wallets, que deben permanecer independientes y seguras frente a amenazas externas. De este modo, se podría lograr un equilibrio eficaz entre medidas de seguridad robustas y las restricciones técnicas propias de estos dispositivos. En conclusión, aunque las hardware wallets no deben, por diseño, realizar verificaciones que dependan de recursos externos, sí podrían beneficiarse enormemente de comprobaciones internas como las propuestas basadas en principios criptográficos sólidos, particularmente en análisis de entropía y control de diversidad en la generación de semillas mnemónicas. Estas medidas simples pero efectivas podrían proteger mejor a los usuarios y prevenir pérdidas potencialmente devastadoras.

VI. Conclusiones

Existe una deficiencia de seguridad importante en muchas hardware wallets debido a la ausencia de pruebas posteriores a la creación o restauración de una billetera. Esta falta de verificación contrasta con las prácticas de seguridad comunes en otras plataformas, como la comprobación de la fortaleza de las contraseñas.

A pesar de que las hardware wallets están diseñadas para operar de forma aislada y no deben depender de recursos externos para la verificación de seguridad, sí podrían beneficiarse enormemente de comprobaciones internas.

El trabajo propone la incorporación de un sistema de verificación automatizado dentro del firmware del dispositivo que evalúe si la configuración de la wallet es segura antes de su sincronización.

Se sugiere un análisis basado en la aleatoriedad, que podría incluir la evaluación de la entropía y el coeficiente de correlación serial de la clave privada generada o restaurada. Un valor bajo de entropía (por debajo de 3.5) o un alto coeficiente de correlación serial (superior a 0.1 en valor absoluto) podría indicar una secuencia binaria inválida.

También se plantea la implementación de mecanismos para mitigar debilidades relacionadas con estándares como BIP-39, como la comprobación de una mínima diversidad en las palabras mnemotécnicas generadas. Por ejemplo, en una semilla de 12 palabras, se podría requerir que al menos 8 sean únicas.

Se considera que esta propuesta inicial de verificación interna funciona bien y es eficiente en términos de cómputo y memoria, lo que permitiría su implementación en cualquier hardware wallet, incluso con recursos limitados.

En resumen, la incorporación de estas comprobaciones internas basadas en principios criptográficos sólidos podría mejorar significativamente la seguridad de las hardware wallets sin comprometer su rendimiento o usabilidad, ayudando a proteger mejor a los usuarios y prevenir pérdidas de fondos.

Las comprobaciones de seguridad propuestas se basan en el análisis de la aleatoriedad de la clave privada mediante la entropía y el coeficiente de correlación serial, así como en la diversidad de las palabras mnemotécnicas en el caso de semillas BIP-39. Si bien estos métodos son eficientes en términos de cómputo y memoria, podrían considerarse relativamente simples y tal vez no detecten todos los tipos de debilidades o vulnerabilidades posibles en una clave privada o semilla.

El trabajo reconoce que las hardware wallets, por diseño, no deben depender de recursos externos para verificar la seguridad, lo que limita el tipo de comprobaciones que se pueden realizar. Esto implica que no se pueden realizar consultas a bases de datos de claves comprometidas o patrones débiles en línea, como sí se hace en otros entornos.

Futuras investigaciones se centran en la exploración e implementación de pruebas de aleatoriedad y seguridad más avanzadas que puedan ser incorporadas directamente en el firmware de las hardware wallets sin requerir recursos externos significativos. Si bien esta propuesta inicial representa una mejora en seguridad sin coste adicional, se podrían utilizar tests estadísticos más sofisticados.

La evaluación del impacto práctico de las comprobaciones propuestas mediante una experimentación más extensa, tal como se esboza en el artículo, también representa una futura línea de investigación crucial para validar la efectividad de las soluciones propuestas en la detección de claves débiles.

Las medidas que proponemos pueden también aplicarse en otro tipo de billeteras, incluyendo las software (como MetaMask), para ofrecer una capa adicional de protección a sus millones de usuarios. En dichos entornos, es viable implementar validaciones contra bases de datos, lo que, como se ha expuesto, no ha sido posible en nuestro caso debido a los estándares de seguridad asociados al uso de hardware wallets.

Nos gustaría que los resultados finales de la propuesta de investigación aquí planteada pudieran incorporarse a todos, o al menos a la mayoría, de los métodos de gestión de criptoactivos existentes en el ecosistema. Para ello, contemplamos la publicación de la totalidad del código, disponible de forma gratuita y bajo un modelo de código abierto.

Agradecimientos

Este trabajo se ha realizado con los fondos del “Plan de Recuperación, Transformación y Resiliencia”, proyecto C084/23 Ada Byron INCIBE-UAH, financiado por la Unión Europea (Next Generation).

Referencias

[1] K. M. Abdellatif y O. Hérivcaux: ”DeepCover DS28C36: A Hardware Vulnerability Identification and Exploitation Using T- Test and Double Laser Fault Injection”, en *2023 Workshop on Fault Detection and Tolerance in Cryptography (FDTC)*, pp. 89-94, 2023.

[2] M. San Pedro, V. Servant y C. Guillemet: ”Practical side-channel attack on a security device”, en *2019 31st International Conference on Microelectronics (ICM)*, pp. 130-133, 2019.

[3] D. Park, M. Choi, G. Kim, D. Bae, H. Kim y S. Hong: ”Stealing Keys From Hardware Wallets: A Single Trace Side-Channel Attack on Elliptic Curve Scalar Multiplication Without Profiling”, en *IEEE Access*, vol. 11, pp. 44578-44589, 2023.

[4] N. Ivanov y Q. Yan: .EthClipper: A Clipboard Meddling Attack on Hardware Wallets with Address Verification Evasion”, en *2021 IEEE Conference on Communications and Network Security (CNS)*, pp. 191-199, 2021.

[5] A. Holmes y W. J. Buchanan: .A framework for live host-based Bitcoin wallet forensics and triage”, en *Forensic Science International: Digital Investigation*, vol. 44, p. 301486, 2023.

[6] I. Hagui, A. Msolli, N. ben Henda, A. Helali, A. Gassoumi, T. P. Nguyen y F. Hassen: .A blockchain-based security system with light cryptography for user authentication security”, en *Multimed Tools Appl*, vol. 83, pp. 52451–52480, 2024.

[7] S. Alrubei, E. Ball y J. Rigelsford: .Adding Hardware Security into IoT-Blockchain Platforms", en *2022 IEEE Latin-American Conference on Communications (LATINCOM)*, pp. 1–6, 2022.

[8] A. J. Cabrera-Gutierrez, E. Castillo, A. Escobar-Molero, J. A. Alvarez-Bermejo, D. P. Morales y L. Parrilla: İntegration of Hardware Security Modules and Permissioned Blockchain in Industrial IoT Networks", en *IEEE Access*, pp. 1-4, 2022.

[9] N. Lehto, K. Halunen, O. -M. Latvala, A. Karinsalo y J. Salonen: ÇryptoVault - A Secure Hardware Wallet for Decentralized Key Management", en *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, pp. 1–4, 2021.

[10] H. Rezaeighaleh y C. C. Zou: "New Secure Approach to Backup Cryptocurrency Wallets", en *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2019.

[11] L. -Y. Yeh, W. -H. Hsu, J. -L. Huang, C. -Y. Shen y C. Wu-Lee: Ïntegrating Cellphone-based Hardware Wallet with Visional Certificate Verification System", en *GLOBECOM 2020*, pp. 1–6, 2020.

[12] R. Kunduru, R. Thatikonda, S. A. Vaddadi, B. Lakshmi, Y. A. D. Isravel y V. Khirasaria: .Applications of Blockchain Technology for Secure Transaction of Electronic Health Records", en *2023 2nd InternationalConference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 201–207, 2023.

[13] D. Paulin, R. Joud, C. Hennebert, P.-A. Moëllic, T. Franco-Rondisson y R. Jayles: "HistoTrust: tracing AI behavior with secure hardware and blockchain technology", en *Ann. Telecommun*, 2023.

[14] R. K. Raman y L. R. Varshney: Çoding for Scalable Blockchains via Dynamic Distributed Storage", en *IEEE/ACM Trans. Netw*, pp. 1–14, 2021.

[15] S. Raveenthiren y S. Peraisoody: .A novel approach to secure IoT cryptocurrency wallets based on Enhanced ECDSA algorithm", en *5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pp. 1–6, 2024.

[16] A. Brighente, M. Camaioni, M. Conti y E. Olivastri: "Should I Mine or Should I Break: On the Worthiness of Brute-Forcing Cryptocurrency Addresses", en *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 279–284, 2022.

[17] A. S. Nair, S. M. Thampi y V. Jafeel: .A Post-quantum Secure PUF based Cross-Domain Authentication Mechanism for Internet of Drones", en *Veh. Commun*, p. 100780, 2024.

[18] Y. Takei y K. Shudo: "Pragmatic Analysis of Key Management for Cryptocurrency Custodians", en *2024 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pp. 747–765, 2024.

[19] T. Thomas, M. Piscitelli, I. Shavrov y I. Baggili: "Memory FORESHADOW: Memory FOREnSics of HArDware CryptOcurrency wallets – A Tool and Visualization Framework", en *Forensic Sci. Int.: Digit. Investigation*, vol. 33, p. 301002, 2020.

Técnica de Fusión Audio-Visual para la Detección de Contenido Sensible en Videos

Daniel Povedano Álvarez, Ana Lucila Sandoval Orozco, Luis Javier García Villalba*

Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA)
Facultad de Informática, Despacho 431, Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid
Email: {dpovedano, asandov, javierv}@ucm.es

***Resumen*—La detección de contenido sensible en videos en línea es crucial para la seguridad digital y la moderación de contenido. Este trabajo presenta Multimodal Audio-Visual Attention (MAV-Att), un modelo multimodal basado en Aprendizaje Profundo que integra señales de audio y video para mejorar la detección. Evaluado en el conjunto de datos LSPD, MAV-Att analiza 52.427 segmentos de 20 segundos con una extracción de fotogramas clave optimizada. Su arquitectura incluye dos ramas paralelas para audio e imagen, potenciadas por mecanismos de atención para modelar dependencias temporales y relaciones multimodales. Entrenado con una función de pérdida combinada, MAV-Att alcanzó una puntuación F1 del 94,9 % en segmentos y 94,5 % en videos completos, superando el estado del arte previo en 6,75 %. Con solo 12 millones de parámetros, es un modelo ligero y eficiente, estableciendo un nuevo referente en la detección de contenido sensible mediante fusión multimodal.**

***Index Terms*—Aprendizaje profundo, Espectrogramas de Mel, Clasificación de vídeos, Análisis de audio, Aprendizaje multimodal, Mecanismos de atención**

Tipo de contribución: *Investigación original (límite 8 páginas)*

I. Introducción

En los últimos años, la proliferación de dispositivos multimedia, como los teléfonos móviles, junto con la difusión masiva de contenido en plataformas de redes sociales, ha provocado un aumento significativo del contenido ilegal en Internet. Este incremento incluye material sensible y sexualmente explícito, lo que representa graves riesgos para los usuarios, especialmente para niños y adolescentes. La facilidad con la que las imágenes y videos ilegales se propagan en plataformas digitales amenaza el bienestar mental de los jóvenes, aumentando su exposición a riesgos psicosociales [1]. Durante la pandemia de COVID-19, estas dinámicas se intensificaron aún más, ya que las medidas de confinamiento y las dificultades económicas contribuyeron al aumento del Abuso Sexual Infantil (ASI) y la Explotación Sexual Infantil (ESI), lo que resultó en consecuencias físicas, psicológicas y emocionales tanto a corto como a largo plazo.

La integración de la Inteligencia Artificial (IA) en los esfuerzos para combatir el crimen ha mostrado resultados prometedores, mejorando significativamente las iniciativas de protección infantil [2].

En este contexto, la identificación y prevención de contenido sensible, como el ASI, se ha convertido en un objetivo clave para las fuerzas del orden y las actividades forenses. Las herramientas automatizadas facilitan la detección de contenido sexualmente explícito, restringiendo su acceso en entornos sensibles (por ejemplo, escuelas y espacios públicos), al tiempo que apoyan los análisis forenses dirigidos a identificar a los perpetradores y víctimas [3]. Sin embargo, esta tarea sigue siendo compleja, ya que el análisis del ASI está estrictamente regulado por la ley, lo que requiere herramientas que operen eficazmente sin acceso directo al material original.

A pesar de los avances, los métodos multimodales que combinan diversas características del contenido—como datos visuales, auditivos y de movimiento—siguen enfrentando desafíos para detectar con precisión el contenido sensible, especialmente cuando el material explícito no es visualmente evidente. Por ejemplo, se espera que las metodologías del Aprendizaje Profundo (AP), incluidas las arquitecturas supervisadas, no supervisadas y auto-supervisadas, evolucionen aún más para optimizar la detección de contenido sensible [4]. Estudios recientes destacan que, aunque la combinación de características visuales, auditivas y textuales mejora la detección, confiar únicamente en el contenido visual presenta limitaciones significativas. Esto es particularmente cierto dada la naturaleza a menudo ambigua del material de ASI, donde el contenido sexual explícito puede no ser siempre visible [5]. Notablemente, tras la creación de un conjunto de datos anotado de imágenes de Material de Abuso Sexual Infantil (MASI), las estadísticas revelaron que casi el 35 % de las muestras no contenían desnudez, lo que podría generar falsos negativos en predicciones de diversos enfoques que dependen únicamente de la detección de contenido sexual basado en imágenes o fotogramas.

Para abordar estas limitaciones, la integración de características de audio ha surgido como una solución prometedora. Técnicas como el análisis de espectrogramas y la transcripción de voz a texto (por ejemplo, usando *Whisper*) han demostrado ser eficaces en la reducción de falsos negativos. El análisis de audio desempeña un papel crucial, especialmente cuando el análisis visual tradicional no es suficiente, ya que puede revelar patrones indicativos de material sensible que de otro modo podrían pasar desapercibidos. Por ejemplo, [6] combinaron características de audio y visuales utilizando mecanismos de atención, logrando una precisión del 93,4 % en la detección de contenido explícito. De manera similar, [7] introdujeron un esquema de apilamiento multimodal que mejoró la precisión de detección al 92,33 % en tareas de monitoreo en línea.

Sin embargo, es importante destacar que estos estudios

recientes no han implementado un mecanismo robusto para capturar de manera efectiva la relación entre las características de audio e imagen. Los enfoques existentes suelen basarse en estrategias heurísticas que simplemente concatenan características de ambas modalidades, sin modelar sus interacciones intrincadas.

Para superar esta deficiencia, la integración de mecanismos de atención y fusión surge como una solución prometedora. Al permitir que el modelo enfoque dinámicamente las interacciones significativas entre las señales auditivas y visuales, puede mejorar significativamente la detección, especialmente en escenarios donde el contenido sensible no es visualmente explícito, pero puede inferirse a partir de la señal de audio. La ausencia de una operación avanzada de este tipo representa una brecha crítica y una oportunidad para futuros avances en los sistemas multimodales de detección de contenido sensible.

En este trabajo se presenta MAV-Att, una arquitectura multimodal basada en mecanismos de atención y fusión jerárquica, diseñada específicamente para abordar estos desafíos en la detección de Material de Abuso Sexual Infantil (MASI). MAV-Att permite capturar de forma explícita las relaciones complejas entre las modalidades visual y auditiva, mejorando la precisión en contextos donde una de las señales es ambigua o incompleta. Esta es la principal contribución del trabajo.

Este trabajo busca contribuir a los esfuerzos en curso para desarrollar sistemas de detección más efectivos para contenido sensible mediante el aprovechamiento de enfoques multimodales que integren tanto características auditivas como visuales. Esta capacidad mejorada de detección puede, a su vez, respaldar la identificación de MASI y MESI, mitigando así el riesgo de falsos negativos y mejorando la precisión general de los sistemas de moderación de contenido.

El artículo está organizado en siete secciones. La Sección I introduce el contexto en el cual se ha desarrollado este trabajo, seguido de la revisión del estado del arte sobre la detección de contenido sensible (sexual) en la Sección II. La Sección III proporciona una descripción detallada del conjunto de datos utilizado en este estudio. La Sección IV detalla la metodología propuesta, mientras que la Sección V presenta los experimentos realizados y la Sección VI discute los resultados correspondientes. Finalmente, la Sección VII resume las conclusiones clave extraídas de la investigación.

II. Trabajos relacionados

En los últimos años, los avances en AP han transformado la detección de contenido sexual, alcanzando niveles sin precedentes de precisión y eficiencia. Estos avances se han logrado mediante el uso de características visuales, auditivas y multimodales (combinando características visuales y acústicas), cada una de las cuales contribuye significativamente a la identificación y clasificación de material de explotación infantil y abuso sexual en línea. Estas técnicas han optimizado la capacidad de detectar contenido ilegal y comportamientos sospechosos, facilitando su prevención y mitigación mediante sistemas automatizados avanzados.

II-A. Características visuales

Esta sección describe los enfoques basados en características visuales utilizando algoritmos de AP. Estos métodos emplean Redes Neuronales Convolucionales (RNC) y otras arquitecturas avanzadas para identificar contenido sensible, partiendo de la premisa de que dicho material a menudo contiene patrones distinguibles, como formas específicas, texturas o distribuciones de colores. Los modelos de AP extraen automáticamente características visuales de alto nivel, lo que permite un reconocimiento más preciso de contenido explícito en comparación con los métodos tradicionales. Al entrenarse en conjuntos de datos anotados, estos algoritmos pueden detectar y clasificar contenido visual sensible, incluidos patrones sutiles y complejos que indican material inapropiado.

En técnicas anteriores [8][9], los umbrales manuales para el reconocimiento de piel y la detección de movimiento limitaban la adaptabilidad y precisión. AP mitigó estos desafíos, permitiendo la extracción automática de características de alto nivel y mejorando la precisión y generalización de los sistemas de detección de contenido sensible.

Desde la aparición de las RNC para resolver tareas de visión por computadora, recientemente se han empleado con éxito en la detección de contenido sensible.

Vitorino et al. [10] se centran en aprovechar las redes neuronales profundas para combatir la pornografía infantil en la era de las redes sociales. Los autores aplican redes neuronales avanzadas para identificar y clasificar automáticamente contenido explícito, especialmente relacionado con la explotación infantil, en plataformas de redes sociales. Este enfoque tiene como objetivo ayudar en la detección de material ilegal a gran escala, ayudando a organizaciones y autoridades a combatir el ASI de manera más efectiva en un entorno digital cada vez más complejo.

Pérez et al. [11] extendieron estos esfuerzos integrando características de movimiento (flujo óptico y vectores) con GoogLeNet y clasificadores MSV (Máquina de Soporte Vectorial, del inglés *Support Vector Machine* (SVM)). Su enfoque logró una precisión del 97,9 % y 96,4 % en los conjuntos de datos *Pornography-800* y *Pornography-2k*, respectivamente.

Mallmann et al. [12] presentaron *Private Parts Censor* (PPCensor), basado en Faster R-CNN, para la detección de pornografía en tiempo real. Entrenado en su conjunto de datos *Private Parts Object*, PPCensor logró un 2,34 % Ratio de Falsos Negativos (RFN) en el conjunto de datos *Pornography-2k*, mejorando los resultados en un modesto 0,31 %.

Aldahoul et al. [13] utilizaron YOLO para la detección de regiones, seguida de RNCs y clasificadores SVM, logrando una exactitud del 95,04 % en su conjunto de datos personalizado. Combinando ResNet50 y YOLO-ResNet101 con Random Forest se obtuvo un puntuación F1 de 90,03 % , demostrando un alto rendimiento en la clasificación de contenido pornográfico.

Gangwar et al. [14] desarrollaron AttM-CNN, una arquitectura RNC saprovechando mecanismos de atención, entrenada en *Pornographic-2M*, logrando una exactitud del 97,10 % en el conjunto de datos *Pornography-2k*, mostrando resultados superiores para la clasificación de pornografía y grupos de edad.

Rautela et al. [15] propusieron ODT, un marco basado en transformers, logrando un 99,42 % de exactitud en *Pornography-2k* y 99,04 % de exactitud en *Pornography-800*, ofreciendo resultados de última generación en detección de obscenidad computacionalmente eficiente.

Rautela et al. [16] introdujeron DVRGNet, un marco

jerárquico de RNCs que incorpora datos de movimiento, logrando un 99,42 % de exactitud en *Pornography-2k* y un 99,04 % de exactitud en *Pornography-800*, con experimentos extensivos que validan su efectividad en la detección de contenido obsceno.

Samal et al. [17] presentaron un enfoque de aprendizaje por transferencia que combina MobileNet V2 y DenseNet169 para la detección de imágenes obscenas, alcanzando una exactitud del 98,50 % en *Pornography-800* y *Pornography-2k*, aunque las lagunas metodológicas limitan su generalización.

Gangwar et al. [18] propusieron DeepHSAR, un marco semi-supervisado para la detección detallada de actividad sexual, logrando una exactitud del 99,85 % en *Pornography-2k* y resultados competitivos en los conjuntos de datos NUS-WIDE y MS-COCO.

II-B. Características auditivas

Mientras que las características visuales han dominado la detección de objetos, las características auditivas están emergiendo como una herramienta complementaria poderosa. Integrar audio puede proporcionar un contexto adicional que no está presente en las imágenes.

Detectar contenido sensible en archivos de audio representa un desafío considerable, y abordarlo exige el uso de técnicas avanzadas de Aprendizaje Automático (AA). Este análisis se centra en explorar e investigar las últimas investigaciones en este campo, prestando especial atención a cómo se están aplicando metodologías de AA para mitigar este problema.

Esta sección revisa las contribuciones más significativas en el campo de la detección de contenido sensible utilizando características acústicas a través de AA y IA.

Lovenia et al. [19] propusieron un método basado en audio para la detección de pornografía, utilizando RNCs entrenadas en espectrogramas *log Mel*. Su enfoque logró puntuaciones F1 del 94,89 % en segmentos de audio (60 segundos) y del 92,02 % a nivel de audio (video) en el conjunto de datos Pornography-800 [20]. Zhou et al. [21] se centraron en un modelo basado en RNC con mejoras como mecanismos de atención, logrando una precisión del 97,19 % en un gran conjunto de datos de 224.127 audios pornográficos y 274.206 audios normales. Liu et al. [22] introdujeron el Dual-Path Fused Transformer Net (DPFTNet), combinando características *log Mel*,*MFCC* y *GFCC*, logrando una exactitud del 93,20 % y una puntuación F1 del 93,56 % en su conjunto de datos de 7.942 muestras de audio pornográfico.

II-C. Características multimodales

La integración de características visuales y auditivas en un marco multimodal está llevando la detección de objetos a nuevas fronteras. Los modelos multimodales pueden procesar y fusionar información de múltiples fuentes, proporcionando una comprensión más completa y robusta del entorno.

Song et al. [23] introdujeron un método que utiliza características multimodales (audio, movimiento, imagen y flujo óptico) para la detección de contenido sensible, logrando una exactitud del 88,3 % para los detectores de movimiento y del 80 % para los basados en audio. Wang et al. [24] se centraron en reconocer pornografía en transmisiones de video en vivo utilizando características multimodales, logrando una exactitud del 74,86 %. Song et al. [7] propusieron un esquema de apilamiento multimodal para la detección en línea, logrando una exactitud del 92,33 %. Fu et al. [6] presentaron PornNet, combinando características auditivas y visuales con mecanismos de atención, logrando una exactitud del 93,4 %. Zhao et al. [25] desarrollaron Porn2Vec, logrando una puntuación F1 del 99,25 % para la detección de sitios web pornográficos. Estudios recientes han explorado el uso de mecanismos de atención cruzada en tareas multimodales, mejorando el rendimiento en el reconocimiento de emociones y la detección de contenido sensible [26], [27], [28].

La fusión de datos visuales y de audio permite a los modelos de AP comprender mejor el contexto y tomar decisiones más informadas. Dado el éxito de la combinación de ambas características, nuestro enfoque busca aprovechar esta integración multimodal para lograr un rendimiento más preciso y robusto.

Además, hemos entrenado nuestro modelo en un conjunto de datos más grande y robusto, llamado LSPD [29]. Este conjunto de datos, el primero de su tipo tanto para tareas de detección de objetos como para clasificación de imágenes y videos en esta área, tiene como objetivo mejorar el estándar de calidad en la clasificación de contenido visual pornográfico y la detección de objetos sexuales. El conjunto LSPD contiene un corpus a gran escala de imágenes y videos tanto pornográficos como no pornográficos, abarcando una amplia diversidad contextual.

En contraste con los modelos mencionados anteriormente, nuestro enfoque emplea una metodología multimodal y es entrenado de extremo a extremo para explotar de manera efectiva tanto las características de audio como las visuales, permitiendo una mejor toma de decisiones y comprensión del contexto.

III. Conjunto de Datos

La preparación del conjunto de datos es crucial para la detección de contenido sensible. Se ha seleccionado el conjunto de datos LSPD [29] debido a sus características únicas, que incluyen etiquetas a nivel de clase y máscaras poligonales para identificar objetos sexuales privados en imágenes y videos. Con un corpus de 500.000 imágenes y 4.000 videos, el conjunto es robusto para el entrenamiento y la evaluación, al estar compuesto por 2.000 videos seguros y 2.000 videos de contenido sexual. La Tabla I resume la cantidad de videos por clase y la disponibilidad de audio (* indica que un vídeo de la clase sexual no pudo extraerse por encontrase dañado).

Tabla I
Estadísticas de los videos en el conjunto de datos LSPD.

Clase	Total de videos	Videos con audio	Videos sin audio
Seguro	2.000	641	1.359
Pornografía	1.999*	1.660	339
Total	3.999	2.301	1.698

IV. Detección de contenido sensible en audio y vídeo

Este trabajo propone **MAV-Att**, un enfoque multimodal novedoso que combina información audiovisual mediante mecanismos de atención para mejorar la fusión de características y el aprendizaje de representaciones. La arquitectura opera de

manera *end-to-end*, optimizando conjuntamente las modalidades de audio y visual mediante retropropagación. Para ello, se integran segmentos de audio y fotogramas clave alineados temporalmente, lo que potencia la comprensión multimodal del contenido.

Una función de pérdida combinada optimiza las ramas auditiva y visual, asegurando contribuciones equilibradas de cada modalidad. El flujo de trabajo se estructura en las siguientes etapas:

1. **Preprocesamiento:**
   - Se segmentan los videos en clips de 20 segundos, siguiendo prácticas de vanguardia en análisis de audio [19].
   - Se extraen fotogramas clave alineados con los segmentos de audio, aplicando técnicas de *perceptual hashing* (pHash) para eliminar redundancias visuales.
2. **Extracción de Características:**
   - Se obtienen características temporales y espectrales del audio transformándolo en espectrogramas *log-mel*.
   - Se extraen características espaciales y semánticas de los fotogramas clave mediante redes neuronales convolucionales.
3. **Atención Unimodal:** Cada modalidad pasa por un mecanismo de auto-atención que refina sus características más relevantes de forma independiente.
4. **Fusión y Optimización:** Las salidas de las ramas de audio y visual se combinan mediante clasificadores independientes y una función *softmax* ponderada para obtener la clasificación final del contenido (seguro o sensible).

El método propuesto integra las modalidades utilizando mecanismos de auto-atención avanzados, lo que permite al modelo centrarse en las características más relevantes y potenciar la interacción entre ambas fuentes de información. La Fig. 1 ilustra el flujo de trabajo general de la arquitectura MAV-Att.

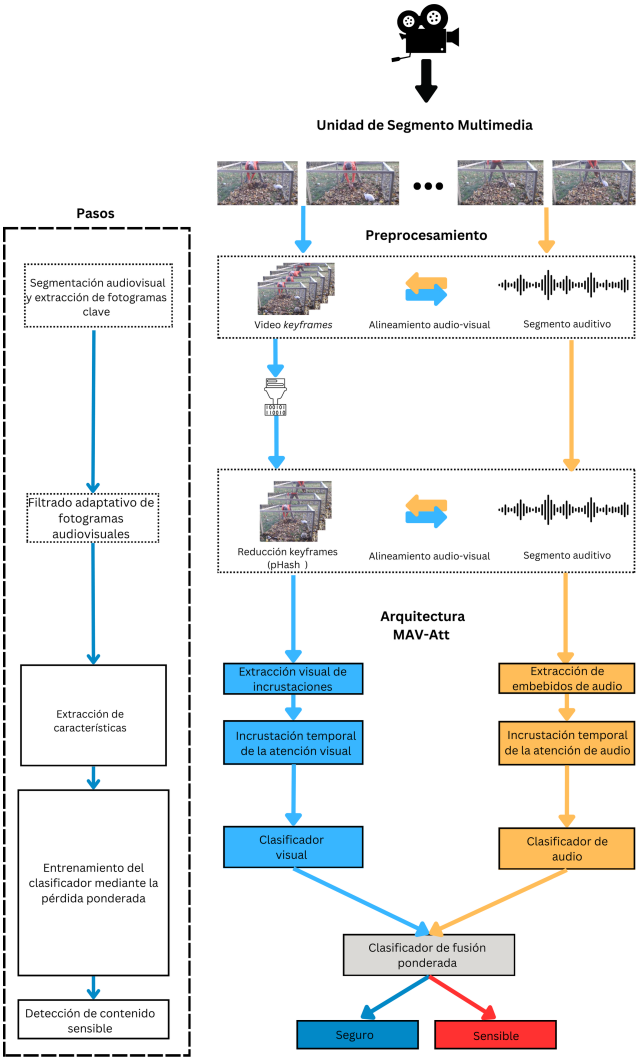


Figura 1. Descripción general de la arquitectura y el flujo de trabajo de MAV-Att.

IV-A. Modelo pre-entrenado usado: ConvNeXt V2

El modelo tConvNeXt V2 -femto- (5,2 millones de parámetros aproximadamente), que sirve como la columna vertebral de este trabajo, se basa en el éxito de los *Vision Transformers* (ViTs) mientras conserva las fortalezas de las RNCs. Moderniza la arquitectura ResNet con optimizaciones clave, como la normalización de capas y las convoluciones eficientes, mejorando tanto el rendimiento como la eficiencia computacional. Este modelo logra resultados de vanguardia mientras se mantiene escalable y adaptable para ajustes finos en una variedad de tareas, incluidas las aplicaciones multimodales.

Para más detalles, consulte el artículo original de Woo et al. [30].

IV-B. Preprocesamiento

Se diseñó un flujo estructurado de preprocesamiento para optimizar el aprovechamiento de las características audiovisuales del conjunto de datos. Este proceso incluyó la limpieza, reducción de ruido y normalización de las entradas, así como la sincronización de fotogramas visuales con sus correspondientes segmentos de audio. Además, se aplicaron técnicas de aumento de datos (como recorte y volteo aleatorio en los fotogramas y cambios en la tonalidad en los segmentos de audio) para diversificar el conjunto de entrenamiento, con el fin de mejorar la capacidad del modelo para generalizar frente a diferentes variaciones de entrada. Este enfoque también permitió aumentar el volumen de datos disponibles sin necesidad de recopilar más muestras, lo que es especialmente útil cuando el acceso a grandes cantidades de datos etiquetados es limitado. Para garantizar la calidad y consistencia de los datos, también se implementaron técnicas de alineación temporal, lo que permitió mejorar la coherencia entre el contenido visual y sonoro de cada segmento. Estas medidas contribuyeron a generar un conjunto de datos más robusto, reduciendo la variabilidad en la entrada y mejorando la capacidad de generalización del modelo. Así, se logró una mayor estabilidad en los resultados durante las fases de entrenamiento y validación, minimizando los errores derivados de inconsistencias temporales o de calidad en los datos.

La extracción de datos se realizó mediante un algoritmo en *Python* utilizando la librería *FFmpeg* [31] y *moviepy* [32]

que segmenta los videos en clips de 20 segundos y alinea dichos segmentos con sus fotogramas clave. Posteriormente, se empleó un algoritmo basado en *pHash* (con una distancia de *Hamming* de 10) para reducir la redundancia visual, estableciendo un mínimo de 2 y un máximo de 10 fotogramas clave para segmentos de 20 segundos. Esta estrategia de segmentación y reducción de redundancia permitió mantener la diversidad en el conjunto de datos sin sacrificar la calidad, lo que resulta crucial para entrenar modelos que puedan adaptarse a una amplia gama de entradas audiovisuales.

Los resultados del preprocesamiento se resumen en la Tabla II:

Tabla II
ESTADÍSTICAS PARA SEGMENTOS DE 20 SEGUNDOS DESPUÉS DEL PREPROCESAMIENTO.

Etiqueta	Estadística	Segmentos (20s)
Seguro	Total de archivos WAV	21.442
	Total de imágenes PNG	136.687
Sexual	Total de archivos WAV	30.985
	Total de imágenes PNG	140.446
General	Total de archivos WAV	52.427
	Total de imágenes PNG	277.133

IV-C. Métricas de rendimiento

El rendimiento se ha expresado en términos puntuación F1 (del inglés *F1-measure*) y de la exactitud (del inglés *accuracy*), priorizando puntuación F1 macro durante la búsqueda de los mejores hiper-parámetros durante la validación cruzada (5 *fold*).

La métrica puntuación F1 macro es comúnmente utilizada en la evaluación de modelos de clasificación, especialmente cuando se enfrentan conjuntos de datos desequilibrados, es decir, cuando algunas clases tienen muchos más ejemplos que otras. De esta manera, se calcula el puntuación F1 para cada clase por separado y luego calcula el promedio no ponderado de estas medidas puntuación F1 individuales. Esto significa que cada clase contribuye de manera igual al resultado final, independientemente de su tamaño o distribución en el conjunto de datos.

Sean los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN), respectivamente, la exactitud de clasificación , puntuación F1 y puntuación F1 macro se definen como sigue:

$$Exactitud = 100 \cdot \frac{VP + TN}{VP + FP + VN + FN} \tag{1}$$

$$Precisión = \frac{VP}{VP + FP} \tag{2}$$

$$Sensibilidad = \frac{VP}{VP + FN} \tag{3}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{Precisión \cdot Sensibilidad}{\beta^2 \cdot Precisión + Sensibilidad} \tag{4}$$

donde β = 1 para puntuación F1.

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{5}$$

V. CONFIGURACIÓN EXPERIMENTAL Y PROTOCOLO DE EVALUACIÓN

Este estudio explora el potencial de una arquitectura multimodal para aprovechar de forma simultánea las características audiovisuales. Se realizaron experimentos utilizando la metodología descrita en la Sección IV-B para seleccionar un modelo robusto, capaz de lograr una comprensión integral de escenas complejas, especialmente en tareas de clasificación desafiantes.

V-A. Configuración de los Experimentos

Se utilizó una validación cruzada de 5 particiones para evaluar el rendimiento de cada modelo, asegurando una evaluación exhaustiva. Se definieron los hiperparámetros y configuraciones del algoritmo, siendo el marco especialmente adecuado para tareas donde las características audiovisuales son críticas (por ejemplo, detección de actividad sexual). Tras la selección del modelo, las predicciones se realizarán mediante votación mayoritaria en cada segmento de audio y fotograma.

V-B. Protocolo de Evaluación

Los experimentos se efectuaron sobre el conjunto de datos anotado de 3.999 videos (conjunto LSPD, Sección III) dividido el conjunto de datos de videos en subconjuntos de entrenamiento y prueba (80/20). La partición se realizó sobre el conjunto de vídeos y no directamente sobre los segmentos audio-visuales para evitar que segmentos del mismo vídeo cayeran en el conjunto de entrenamiento y de prueba (fuga de datos). Esto provocaría que se hubieran obtenido métricas sobre un conjunto de prueba no verdaderamente independiente, generando una sobrestimación del rendimiento del modelo. Durante la validación se utilizaron el puntuación F1 y la precisión, mientras que en la prueba se evaluaron además precisión, sensibilidad y la matriz de confusión.

La matriz de confusión se define con:

- VP: Instancias positivas correctamente predichas.
- VN: Instancias negativas correctamente predichas.
- FP: Instancias negativas predichas como positivas.
- FN: Instancias positivas predichas como negativas.

V-C. Detalles de la Implementación

La implementación se realizó en Python 3.11.7, ejecutándose en un procesador Intel® Core™ i9-13900 (2.00 GHz), 32 GB de RAM y una NVIDIA RTX 3050 con 8 GB de memoria. Se configuró una tasa de aprendizaje de 1×10^{-4} con el optimizador AdamW (*weight decay* de 1×10^{-3}), utilizando un mini-batch de 32 y entrenando por 50 épocas con *early stopping* (paciencia de 5).

Para el preprocesamiento:

- Fotogramas: Las imágenes se redimensionaron a 224×224 mediante interpolación bicúbica y se normalizaron usando la media y desviación estándar de ImageNet (conjunto de datos con el que se suelen pre-entrenar los modelos de AP para tareas de Visión artificial).
- Audio: Se convirtieron los datos en espectrogramas *log-mel* utilizando los siguientes parámetros:
  - $N_{\text{FFT}} = 1024$
  - Frecuencia de muestreo: 16.000 Hz.

- $N_{\text{MELS}} = 224$
- Hop length: 512.

Estos espectrogramas se normalizaron siguiendo la media y desviación típica esperadapor los modelos preentrenados en ImageNet [33].

Además, se aplicaron técnicas de aumento de datos en ambas modalidades para aumentar la variabilidad y robustez del modelo.

En el procesamiento en *batch* (lote), se desempaquetaron los distintos componentes de cada lote: audios, imágenes, longitudes de fotogramas clave por segmento y las etiquetas (seguro/sensible (sexual)). Las imágenes se concatenaron en un tensor único (agrupando todos los fotogramas), las longitudes se convirtieron en tensores para preservar la secuencia, y los audios se apilaron en un tensor (asegurando la correspondencia entre muestras y longitudes).

V-D. Experimento con Segmentos de 20 Segundos

Para evaluar el rendimiento del modelo se utilizaron segmentos audiovisuales de 20 segundos extraídos de un conjunto de datos diverso. En la fase inicial se realizaron experimentos base sin mecanismos de auto-atención ni aumento de datos para establecer métricas de referencia y realizar la configuración de la función de pérdida combinada. Esto permitió analizar el impacto de las modalidades de audio, imagen y su fusión en la clasificación.

V-D1. Modelos Base para Segmentos de 20 Segundos: Se llevaron a cabo seis experimentos (ver Tabla III) para evaluar diferentes combinaciones de características. Los resultados mostraron como la integración de audio y fotogramas (características visuales) mejora significativamente el rendimiento (comparado con el exprimento 1 y 2, alcanzándose un puntuación F1 máximo del 92,28 % en la configuración óptima.

Tabla III
RESUMEN DE LOS EXPERIMENTOS BASE CON DIFERENTES COMBINACIONES DE CARACTERÍSTICAS Y SUS RESPECTIVAS PUNCUACIONES F1 PROMEDIO TRAS VALIDACIÓN CRUZADA DE 5 PARTICIONES (*5-kfold*).

Exp.	Pesos (pérdida combinada) (audio|imagen|fusión)	Puntuación F1 (%)
Audio	-	80,8
Visual	-	91,03
Audio + Visual + Fusión (Pérdida fusión)	-	92,18
Audio + Visual + Fusión (Pérdida combinada)	1|1|1	92,04
Audio + Visual + Fusión (Pérdida combinada)	0,1|0,2|0,4	92,26
Audio + Visual + Fusión (Pérdida combinada)	0,1|0,3|0,5	92,28

V-D2. Análisis del Rendimiento de MAV-Att: El clasificador final MAV-Att se evaluó utilizando un mecanismo de auto-atención configurado con 6 cabezas a nivel de fotogramas, 4 para audio y 2 para atención cruzada. Se realizaron experimentos adicionales con diferentes configuraciones de capas entrenables (permitir entrenar un número de capas del modelo pr-entrenado seleccionado) y aumento de datos. La Tabla IV resume estos experimentos, donde el Experimento 3 (10 capas entrenables de audio, 15 de imagen y aumento de datos) alcanzó un puntuación F1 promedio del 94,21 %.

Tabla IV
RESUMEN DE EXPERIMENTOS CON DISTINTAS CONFIGURACIONES DE CAPAS ENTRENABLES Y AUMENTO DE DATOS PARA SEGMENTOS DE 20 SEGUNDOS.

Exp.	Capas entrenables de Audio	Capas entrenables de Imagen	Aumento de datos	Puntuación F1 (%)
1	0	0	No	92,87
2	15	10	Sí	93,86
3	10	15	Sí	94,21

La Figura 2 ilustra, mediante un diagrama de caja, cómo los modelos de los Experimentos 2 y 3 muestran mejoras en la media y en la varianza del desempeño.

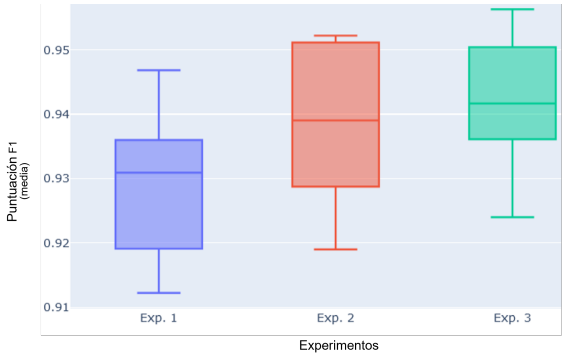


Figura 2. Diagrama de caja de la validación cruzada para segmentos de 20 segundos.

V-D3. Evaluación en el Conjunto de Prueba: La evaluación en el conjunto de prueba, con segmentos de 20 segundos, evidenció un puntuación F1 del 94,8 % y una exactitud del 94,9 %, lo que confirma la capacidad del modelo para generalizar a datos no vistos. La Tabla V presenta un informe detallado de la clasificación, con métricas clave como la precisión, la sensibilidad, la puntuación F1 y el soporte para cada clase, incluyendo las etiquetas Seguro y Sensible. Se observa que ambas clases tienen un rendimiento excelente, con la etiqueta 'Sensible' mostrando una precisión ligeramente superior (95,1) en comparación con la etiqueta Seguro (94,7). La puntuación F1 para ambas clases también es alta, lo que indica un equilibrio adecuado entre precisión y sensibilidad.

Además, se proporcionan los valores de las medias globales, como la precisión global (94,9 %) y las medias macro y ponderada, que refuerzan la robustez del modelo en su conjunto. Estos resultados sugieren un buen desempeño en la clasificación general, con una adecuada capacidad para identificar correctamente las instancias de cada clase.

Por otro lado, la Figura 3 muestra la matriz de confusión, que complementa este análisis al destacar la baja incidencia de falsos positivos y negativos. Esto refuerza la efectividad del modelo, indicando que las predicciones son mayormente correctas, lo cual es crucial para aplicaciones donde la precisión y la fiabilidad son esenciales.

En conjunto, estos resultados evidencian el alto rendimiento del modelo (teniendo en cuanta que es un modelo con pocos parámetros), que ha sido capaz de generalizar correctamente a partir de los datos de prueba.

Tabla V
INFORME DE CLASIFICACIÓN CON PRECISIÓN, SENSIBILIDAD, PUNTUACIÓN F1 Y SOPORTE PARA EL CONJUNTO DE PRUEBA.

Etiqueta	Precisión (%)	Sensibilidad (%)	Puntuación F1 (%)	Soporte
Seguro	94,7	93,58	94,1	4627
Sensible	95,1	95,98	95,6	6048
Exactitud global			94,9	10675
Media Macro	94,9	94,8	94,8	10675
Media Ponderada	94,9	94,9	94,9	10675

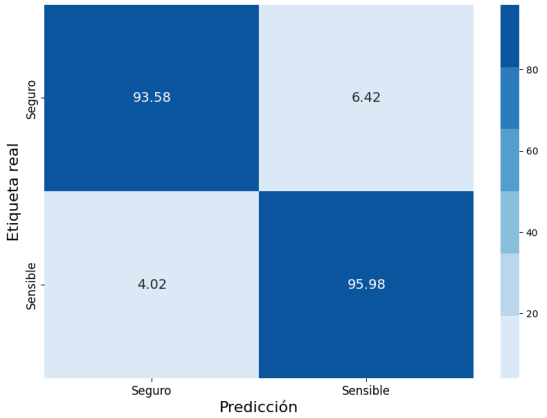


Figura 3. Matriz de confusión para el conjunto de prueba con segmentos de 20 segundos.

VI. RESULTADOS FINALES

A continuación se presentan los resultados finales para segmentos de 20 segundos, así como para videos completos, donde se ha utilizado la votación mayoritaria para la agregación de predicciones segmentadas. Este método consiste en asignar a cada video la clase más frecuente (moda) entre sus segmentos, permitiendo obtener una decisión global a partir de múltiples predicciones parciales.

Nuestra metodología supera enfoques previos en clasificación de contenido multimodal, mejorando tanto la precisión como la especificidad por clase. La combinación de fusión audiovisual y mecanismos de atención permite capturar información contextual rica, clave para la detección de contenido sensible.

VI-A. Estudio de ablación con segmentos de 20 segundos

La Tabla VI presenta un estudio de ablación que evalúa el impacto de los distintos componentes de la arquitectura en la precisión media y la puntuación F1. Se comparan dos enfoques: MAV, que no incorpora mecanismos de atención, y MAV-Att, que sí los incluye. Los resultados indican que la introducción de mecanismos de atención mejora ambas métricas, con un incremento de un 0,46 % en la puntuación F1. Esta mejora sugiere que la atención contribuye a una mejor representación de las características relevantes, permitiendo al modelo tomar decisiones más precisas. La diferencia, aunque sutil, demuestra la eficacia de la atención en la tarea de clasificación, optimizando la capacidad del modelo para identificar patrones significativos en los datos audiovisuales.

VI-B. Resutados sobre videos con segmentos de 20 segundos

Los resultados en la Tabla VII muestran una puntuación F1 de 94,5 % para la clase segura y 95,48 % para la clase

Tabla VI
ESTUDIO DE ABLACIÓN QUE MUESTRA EL IMPACTO DE LOS DISTINTOS COMPONENTES DE LA ARQUITECTURA EN LA PRECISIÓN MEDIA Y LA PUNTUACIÓN F1 EN EL CONJUNTO DE PRUEBAS.

Enfoque	Segmento	Atención	Exactitud (%)	Puntuación F1 (%)
MAV	20	No	94,44	94,34
MAV-Att	20	Si	94,9	94,8

sensible, lo que sugiere que el uso de segmentos, junto con su correspondiente información visual, y con mecanismos de atención, proporciona un mayor contexto, permitiendo una mejor diferenciación entre clases. Además, la matriz de confusión en la Figura 4 confirma la capacidad del modelo para distinguir entre clases con alta fiabilidad, lo que sugiere su idoneidad para aplicaciones donde la precisión en la detección es crítica.

Tabla VII
INFORME DE CLASIFICACIÓN DE VÍDEOS CON PRECISIÓN, SENSIBILIDAD, PUNTUACIÓN F1 Y SOPORTE UTILIZANDO SEGMENTOS DE 20 SEGUNDOS.

Etiqueta	Precisión (%)	Sensibilidad (%)	Puntuación F1 (%)	Soporte
Seguro	95,4	93,5	94,4	400
Sensible	93,6	95,5	94,5	399
Precisión global			94,5	799
Media Macro	94,5	94,5	94,5	799
Media Ponderada	94,5	94,5	94,5	799

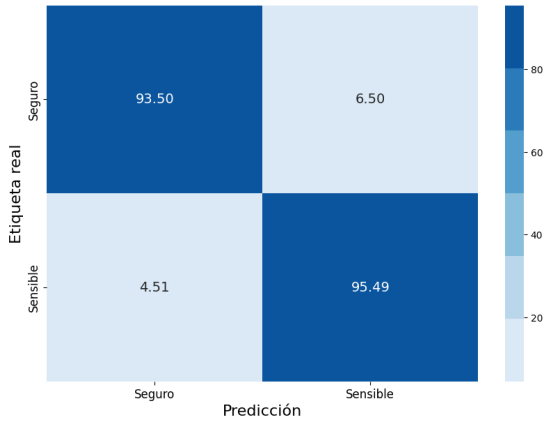


Figura 4. Matriz de confusión de vídeo utilizando el mejor modelo de segmentos de 20 segundos en el conjunto de pruebas.

VII. CONCLUSIÓN Y TRABAJO FUTURO

Este estudio presenta un modelo de clasificación de videos que integra atención visual, atención a características audiovisuales y mecanismos de atención para mejorar la detección de contenido sensible.

Las pruebas muestran que, con segmentos de 20 segundos junto con una selección adecuada de fotogramas clave, el modelo alcanza una puntuación de F1 del 94,1 % para contenido seguro y 95,6 % para contenido sensible, con una puntuación F1 global del 94,8 %. Esto confirma que los segmentos con mecanismos de atención proporcionan información contextual valiosa.

Como limitaciones, la dependencia de ambas modalidades puede afectar el rendimiento cuando una de ellas está ausente. Futuros trabajos explorarán arquitecturas adaptativas,

diferente duración de los segmentos de audio (y numero de fotogramas clave asociados), estrategias de fusión avanzadas y mecanismos de atención

En conclusión, la combinación de modalidades y mecanismos de atención mejora significativamente la clasificación de videos y de segmentos audio-visuales, ofreciendo una solución escalable para la moderación automatizada de contenido digital.

Agradecimientos

Este trabajo se ha realizado con los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation), a través de la Cátedra "Ciberseguridad para la Innovación y la Protección Digital" INCIBE-UCM. Además, este trabajo ha contado con el apoyo del proyecto ALUNA (https://aluna-isf.eu/). Este proyecto ha recibido financiación del programa de investigación e innovación Horizonte 2020 de la Unión Europea en virtud del acuerdo de subvención nº 101084929. El contenido de este artículo no refleja la opinión oficial de la Unión Europea. Los autores son los únicos responsables de la información y las opiniones expresadas en él.



Referencias

[1] S. Ramaswamy and S. Seshadri, "Children on the Brink: Risks for Child Protection, Sexual Abuse, and Related Mental Health Problems in the COVID-19 Pandemic," *Indian Journal of Psychiatry*, vol. 62, no. Suppl 3, p. S404, sep 2020.

[2] The Times, "Ai in crime fighting: Transforming child protection efforts," 2025, accessed: January 26, 2025. [Online]. Available: https://www.thetimes.co.uk/article/ai-there-youre-nicked-tech-is-reshaping-how-we-fight-crime-jb9bv7qh3

[3] J. Cifuentes, A. L. Sandoval Orozco, and L. J. García Villalba, "A Survey of Artificial Intelligence Strategies for Automatic Detection of Sexually Explicit Videos," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3205–3222, 2022.

[4] D. Povedano Álvarez, A. L. Sandoval Orozco, J. P. García-Miguel, and L. J. García Villalba, "Learning Strategies for Sensitive Content Detection," *Electronics*, vol. 12, no. 11, p. 2496, 2023.

[5] C. Laranjeira, J. Macedo, S. Avila, and J. A. dos Santos, "Seeing without Looking: Analysis Pipeline for Child Sexual Abuse Datasets," 2022. [Online]. Available: https://arxiv.org/abs/2204.14110

[6] Z. Fu, J. Li, G. Chen, T. Yu, and T. Deng, "PornNet: A Unified Deep Architecture for Pornographic Video Recognition," *Applied Sciences*, vol. 11, no. 7, p. 3066, 2021.

[7] K. Song and Y. S. Kim, "An Enhanced Multimodal Stacking Scheme for Online Pornographic Content Detection," *Applied Sciences (Switzerland)*, vol. 10, 4 2020.

[8] M. D. C. Polastro and P. M. Da Silva Eleuterio, "NuDetective: A Forensic Tool to Help Combat Child Pornography through Automatic Nudity Detection," *Proc. - 21st Int. Work. Database Expert Syst. Appl. DEXA 2010*, pp. 349–353, 2010.

[9] R. Ap-apid, "An Algorithm for Nudity Detection," in *Proceedings of the 5th Philippine Computing Science Congress*, March 2005, pp. 201–205.

[10] P. Vitorino, S. Avila, M. Perez, and A. Rocha, "Leveraging Deep Neural Networks to Fight Child Pornography in the Age of Social Media," *Journal of Visual Communication and Image Representation*, vol. 50, no. May 2017, pp. 303–313, 2018.

[11] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Video Pornography Detection through Deep Learning Techniques and Motion Information," *Neurocomputing*, vol. 230, no. March, pp. 279–293, 2017.

[12] J. Mallmann, A. O. Santin, E. K. Viegas, R. R. dos Santos, and J. Geremias, "PPCensor: Architecture for Real-time Pornography Detection in Video Streaming," *Future Generation Computer Systems*, vol. 112, pp. 945–955, 2020.

[13] N. Aldahoul, H. A. Karim, M. H. L. Abdullah, M. F. A. Fauzi, A. S. Ba Wazir, S. Mansor, and J. See, "Transfer Detection of Yolo to Focus Cnn's Attention on Nude Regions for Adult Content Detection," *Symmetry (Basel).*, vol. 13, no. 1, pp. 1–26, 2021.

[14] A. Gangwar, V. González-Castro, E. Alegre, and E. Fidalgo, "AttM-CNN: Attention and Metric Learning based CNN for Pornography, Age and Child Sexual Abuse (CSA) Detection in Images," *Neurocomputing*, vol. 445, pp. 81–104, 2021.

[15] K. Rautela, D. Sharma, V. Kumar, and D. Kumar, "Obscenity Detection Transformer for Detecting Inappropriate Contents from Videos," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 10 799–10 814, 2024.

[16] ——, "DVRGNet: An Efficient Network for Extracting Obscenity from Multimedia Content," *Multimedia Tools and Applications*, vol. 83, pp. 28 807–28 825, 3 2024.

[17] S. Samal, R. Nayak, S. Jena, and B. K. Balabantaray, "Obscene Image Detection Using Transfer Learning and Feature Fusion," *Multimedia Tools and Applications*, vol. 82, no. 19, pp. 28 739–28 767, 2023.

[18] A. Gangwar, V. González-Castro, E. Alegre, E. Fidalgo, and A. Martínez-Mendoza, "DeepHSAR: Semi-supervised Fine-grained Learning for Multi-label Human Sexual Activity Recognition," *Information Processing and Management*, vol. 61, 9 2024.

[19] H. Lovenia, D. P. Lestari, and R. Frieske, "What Did i Just Hear? Detecting Pornographic Sounds in Adult Videos Using Neural Networks," in *Proceedings of the 17th International audio mostly Conference*. Association for Computing Machinery, 9 2022, pp. 92–95.

[20] S. Avila, N. Thome, M. Cord, E. Valle, and A. De A. Araújo, "Pooling in Image Representation: The Visual Codeword Point of View," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[21] L. Zhou, K. Wei, Y. Li, Y. Hao, W. Yang, and H. Zhu, "Acoustic Pornography Recognition Using Convolutional Neural Networks and Bag of Refinements," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 840–845.

[22] S. Liu, R. Li, Q. Li, and J. Zhao, "Porn Streamer Audio Recognition based on Deep Learning and Random Forest," *Applied Intelligence*, vol. 53, no. 15, p. 18857–18867, feb 2023. [Online]. Available: https://doi.org/10.1007/s10489-023-04491-x

[23] K. H. Song and Y. S. Kim, "Pornographic Video Detection Scheme Using Multimodal Features," pp. 1174–1182, 2018.

[24] L. Wang, J. Zhang, Q. Tian, C. Li, and L. Zhuo, "Porn Streamer Recognition in Live Video Streaming via Attention-Gated Multimodal Deep Features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 4876–4886, 12 2020.

[25] J. Zhao, M. Shao, H. Peng, H. Wang, B. Li, and X. Liu, "Porn2Vec: A Robust Framework for Detecting Pornographic Websites based on Contrastive Learning," *Knowledge-Based Systems*, vol. 228, 9 2021.

[26] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal Emotion Recognition using Cross Modal Audio-Video Fusion with Attention and Deep Metric Learning," *Image and Vision Computing*, vol. 133, p. 104676, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885623000501

[27] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: Multimodal Speech Emotion Recognition using Cross-attention with Deep Fusion," *Expert Systems with Applications*, vol. 245, p. 122946, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423034486

[28] M. Brousmiche, J. Rouat, and S. Dupont, "Multimodal Attentive Fusion Network for audio-visual event recognition," *Information Fusion*, vol. 85, pp. 52–59, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522000264

[29] D. D. Phan, T. T. Nguyen, K. N. K. Nguyen, Q. H. Nguyen, D. L. Vu, and H. L. Tran, "LSPD: A Large-Scale Pornographic Dataset for Detection and Classification," *International Journal of Intelligent Engineering and Systems*, vol. 15, pp. 198–213, 2022.

[30] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," 2023. [Online]. Available: https://arxiv.org/abs/2301.00808

[31] F. Developers, "Ffmpeg," 2000.

[32] R. Zulko and M. Contributors, "Moviepy - video editing with python," 2012.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

Recreación de ceguera inducida por un ciberataque neuronal en una topología neuronal realista

Victoria Magdalena López Madejska *, Sergio López Bernal *, Gregorio Martínez Pérez*, Alberto Huertas Celdrán*
*Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, 30100 Murcia, España
{victoriamagdalena.lopezm, slopez, gregorio, alberto.huertas}@um.es

***Resumen*—Las interfaces cerebro-máquina (BCIs) permiten la interacción con el cerebro para la adquisición de señales cerebrales y la estimulación neuronal. Sin embargo, estas tecnologías pueden ser vulnerables a ciberataques neuronales, afectando funciones cognitivas clave. Este estudio analiza el impacto de Neuronal Jamming (JAM) en una topología neuronal realista de 230.000 neuronas de la corteza visual primaria (V1) del ratón, con el objetivo de inducir una condición similar a la ceguera. Se estudian varios instantes de ataque sobre diferentes porcentajes de neuronas excitadoras de la capa cuatro de V1. Los resultados muestran que el impacto del ciberataque depende de la actividad espontánea, siendo la ventana de ataque 500-900 ms la más próxima a la ceguera. Además, atacando solo al 5 % de las neuronas objetivo se logra un impacto significativo, con un efecto que puede propagarse entre 700 y 1.000 ms.**

***Index Terms*—Interfaces Cerebro-Máquina, Ciberseguridad, Integridad física, Ciberataques neuronales, Funciones cognitivas, Neurociencia**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Las interfaces cerebro-máquina, o BCIs, (del inglés, Brain-computer interfaces) son dispositivos bidireccionales que interactúan directamente con el cerebro, capaces de adquirir señales cerebrales producidas por el usuario y estimular e inhibir la actividad neuronal. Estos dispositivos se clasifican como invasivos, si es necesario una cirugía para implantar los electrodos, o no invasivos, si se utiliza electrodos sobre el cuero cabelludo. Las primeras son comúnmente usadas en entornos médicos, concretamente para el tratamientos de enfermedades neurodegenerativas como la enfermedad de Parkinson, mientras que las segundas, aunque han sido típicamente utilizadas en el campo de la medicina, han ganado popularidad en escenarios de entretenimiento y videojuegos. Por ello, empresas como Neuralink [1] están centradas en investigar interfaces implantables en miniatura con el objetivo de democratizar el acceso a estos dispositivos.

A pesar de las ventajas que ofrecen las BCIs, no están exentas de problemas de seguridad y privacidad. Además, la reducción de costes de fabricación provoca un incremento de vulnerabilidades en estos dispositivos. Por esta razón, la literatura se ha centrado en documentar problemas relacionados con la integridad [2], la confidencialidad [3] y la disponibilidad de los dispositivos y los datos que gestionan [2], así como con la seguridad física de las personas [4]. Sin embargo, estos trabajos tienen limitaciones, pues abordan solo aspectos muy puntuales de cada una de estas dimensiones de impacto. obvian otras preocupaciones dentro de las BCIs. Por ello, la literatura reciente [5] estudió, de forma exhaustiva, el estado de la ciberseguridad de estas tecnologías, identificando una oportunidad por parte de los atacantes para afectar al correcto funcionamiento del cerebro. En base a estas limitaciones, López Bernal et al. [6] propusieron una nueva familia de ciberataques llamada ciberataques neuronales, capaces de estimular o inhibir la actividad neuronal de forma individual. En estos ataques se asume que el atacante ha tomado el control de la BCI tras explotar vulnerabilidades previamente existentes en el dispositivo. Concretamente, López Bernal et al. [7] presentaron una taxonomía de ocho ciberataques inspirados en ataques conocidos en el área de la seguridad de comunicaciones, alterando el comportamiento espontáneo de las neuronas. Estos ciberataques han sido validados en una topología no realista de 276 neuronas, tras entrenar una red neuronal convolucional (CNN, del inglés, Convolutional Neural Network) simulando la corteza visual del ratón. Adicionalmente, los autores definen métricas para cuantificar el impacto causado por estos ataques sobre la actividad neuronal. Pese a estos grandes avances, estas investigaciones se veían limitadas por la falta de realismo en el modelo neuronal empleado.

Por ese motivo, López Madejska et al. [8] implementaron dos de los ochos ciberataques neuronales sobre una topología neuronal realista compuesta por 450 neuronas. Dicho modelo consiste en una reconstrucción detallada de la corteza visual primaria del ratón, particularmente la capa cuatro, utilizando valores estáticos como entrada visual. También, los autores utilizan métricas de impacto definidas en los trabajos mencionados previamente. Aunque este trabajo supuso un avance sustancial para evaluar el impacto de estos ciberataques, todavía tenía limitaciones en relación a la topología, el estímulo visual usado y la falta de estudios sobre el impacto de estos ataques en funciones cognitivas. Por ello, López Madejska et al. [9] utilizó una reconstrucción completa de la corteza visual primaria (V1), formada por alrededor de 230,000 neuronas junto a tres estímulos visuales realistas. Al igual que en los anteriores estudios, utilizan métricas ya definidas para medir el impacto de los ciberataques implementados.

No obstante, la literatura sigue presentando ciertas limitaciones. Inicialmente, no hay estudios sobre el análisis del impacto de ataques diferenciados por capas de V1, por población y por comportamiento de las neuronas (excitadoras o inhibidoras). Seguidamente, todavía resulta necesario un análisis más exhaustivo sobre el efecto de los ataques en funciones cognitivas críticas, específicamente en el ámbito de la visión. Finalmente, no se han explorado ventanas de ataque superiores a 250 ms ni el impacto que una mayor

duración podría tener en la propagación de un ataque. En base a estas limitaciones, este trabajo trata de ofrecer nuevas orientaciones sobre el efecto de estos ciberataques dentro de la visión, presentando las siguientes contribuciones:

- Inducir una condición temporal de ceguera mediante la aplicación de un ciberataque inspirado en jamming, evaluando su impacto en diferentes instantes usando una ventana temporal de 400 ms y sobre diferentes números de neuronas. Para ello, se implementa Neuronal Jamming (JAM), definido previamente en la literatura [10], en una topología neuronal basada en una reconstrucción de V1 de un ratón. El modelo consta de aproximadamente 230,000 neuronas distribuidas por las seis capas de V1. Las neuronas objetivo fueron seleccionadas de la capa cuatro, concretamente neuronas excitadoras, debido a que son las principales receptoras de información visual por parte del núcleo geniculado lateral (LGN, del inglés Lateral Geniculate Nucleus). Asimismo, este estudio utiliza un estímulo visual realista, concretamente una película monocromática de varios fotogramas con una duración de tres segundos, que reproduce dinámicas naturales del entorno visual. Además, está procesada por los filtros del LGN. Por otro lado, este trabajo emplea dos herramientas principales, Brain Modeling Toolkit (BMTK) [11] y NEST [12], para construir y simular el modelo neuronal.
- Estudiar la similitud entre el impacto causado por diferentes configuraciones de JAM y la ceguera. Para ello, se evalúan los resultados mediante la métrica del número de spikes, definida en la literatura, la cual mide el número de activaciones (potenciales de acción o spikes) de todas las neuronas de la topología en un intervalo especifico. Este estudio demuestra que el impacto de JAM varía en función de la actividad espontánea de los intervalos en los que se ejecuta el ataque, siendo la ventana de 500-900 ms la que más se aproxima a la condición de ceguera. Por otro lado, los resultados indican que atacando un pequeño porcentaje de las neuronas excitadoras de la capa cuatro se genera un impacto significativo en la actividad global, demostrando la importancia de dicha capa en el procesamiento visual. Respecto a la propagación del efecto del ciberataque, se observa que las neuronas tardan entre siete y diez intervalos (700-1,000 ms) en recuperar su actividad espontánea.
- Comparar los resultados con los existentes en la literatura. Aunque la literatura más reciente como este trabajo utilizan la misma topología con el mismo número de neuronas, los enfoques difieren notablemente. Mientras que en la literatura se analizan dos ciberataques neuronales y tres estímulos visuales, este trabajo se centra en inducir un estado de ceguera, realizando un análisis más profundo y específico del impacto de un único ciberataque (JAM) aplicado sobre un estímulo visual realista. Por otra parte, en la literatura, JAM se ejecuta durante 100 ms, mientras que en este estudio se evalúa su efecto en una ventana más amplia de 400 ms. Asimismo, aunque la literatura analiza el impacto del ataque sobre el 25 % y 50 % de las neuronas de la topología, llega a la misma conclusión que este trabajo: no es necesario atacar un gran número de neuronas para alcanzar un impacto notorio. Por último, en cuanto la propagación del ataque, en la literatura es de entre seis y siete intervalos (600-700 ms), mientras que en este estudio se puede propagar hasta diez intervalos (1,000 ms).

Este artículo está organizado de la siguiente forma. La Sección II documenta brevemente el estado del arte sobre ciberataques neuronales en topologías neuronales realistas. Seguidamente, la Sección III detalla el diseño de la solución propuesta, sus componentes fundamentales y la configuración del ciberataque implementado. Posteriormente, la Sección IV muestra los resultados obtenidos de cada experimento, en base al número de spikes. En la Sección V se comparan los resultados obtenidos y, después, con los presentados en la literatura. Finalmente, la Sección VI presenta las conclusiones y trabajo futuro.

II. Estado del arte

La literatura que aborda la ciberseguridad de las BCIs se ha centrado únicamente en aspectos muy concretos, como la posibilidad de suplantar señales cerebrales previamente recopiladas [2], acceder a datos cerebrales sin restricciones [3] o estudiar efectos secundarios derivados de la neuromodulación en el cerebro, el posible daño en el tejido cerebral o el efecto rebote al sobreestimular el cerebro del usuario [4]. Sin embargo, estos estudios son limitados y no ofrecen un análisis detallado sobre la ciberseguridad de las BCIs.

Por ello, López Bernal et al. [5] analizaron de forma exhaustiva el estado de la ciberseguridad de las BCIs para ofrecer una visión estandarizada de esta temática. Los autores unificaron el ciclo de vida de la BCI y ofrecieron un estudio de los posibles ataques, el impacto producido y contramedidas aplicables a cada etapa del ciclo propuesto. Además, destacaron las limitaciones de la literatura y los retos futuros para mejorar la seguridad de estas interfaces. Por otro lado, los mismos autores [6] estudiaron las vulnerabilidades del diseño de los sistemas de neuroestimulación de nueva generación (e.g., Neuralink [1]), desarrollando una nueva familia de amenazas denominada ciberataques neuronales. Estos ataques sobreestimulan o inhiben de forma individual las neuronas accesibles por estas BCIs todavía en etapas de desarrollo incipientes, siguiendo estrategias diferentes. Los autores propusieron Neuronal Flooding (FLO) y Neuronal Scanning (SCA), siendo FLO capaz de sobreestimular un conjunto aleatorio de neuronas en un instante de tiempo, mientras que SCA estimula neuronas individuales de forma secuencial a lo largo de un periodo de tiempo. Debido a la falta de modelos neuronales realistas en el momento de la investigación, los autores emplearon un modelo derivado del entrenamiento de una red neuronal convolucional (CNN, del inglés, Convolutional Neural Network).

Dicho modelo fue transformado en una representación biológica simulada, donde las conexiones y pesos sinápticos proceden de la CNN. Asimismo, la topología está compuesta por 276 neuronas, dividas en tres capas: 200 neuronas en la primera capa, 72 neuronas en la segunda capa y cuatro en la última. El motivo de utilizar una CNN viene dado por las similitudes a nivel estructural y funcionamiento con una red biológica de la corteza visual, indicado por Kuzovkin et al. [13]. La CNN se entrenó para resolver el problema de

un ratón que debe encontrar la salida de un laberinto. Como entrada, la red recibe un estímulo visual simplificado, basado en la disposición de las posiciones cercanas a la ubicación actual del ratón. La simulación tiene una duración de 27 segundos, ya que el camino óptimo para salir del laberinto tiene 27 posiciones, representando que el ratón permanece un segundo en cada posición. Por otro lado, los autores definieron diferentes métricas para poder medir el impacto generado por estos ciberataques: número de spikes, porcentaje de shifts y dispersión temporal. La primera métrica cuantifica el numero de activaciones (potenciales de acción o spikes) de las neuronas atacadas respecto al caso espontáneo. La segunda métrica analiza si los spikes se retrasan o se adelantan respecto a la actividad espontánea y, por último, la dispersión temporal estudia el porcentaje de instantes que han tenido spikes. Finalmente, la investigación mencionada concluye que SCA genera una mayor reducción del número de spikes comparado con FLO, mientras que FLO produce un porcentaje mayor de shifts que SCA. Sin embargo, ambos producen valores similares respecto a la dispersión temporal.

Posteriormente, López Bernal et al. [10] presentaron un nuevo ciberataque llamado Neuronal Jamming (JAM), capaz de inhibir la actividad neuronal durante un periodo de tiempo. Los autores utilizaron la misma topología de 276 neuronas mencionada en el anterior trabajo. Asimismo, utilizaron como métricas de impacto el número de spikes y la dispersión temporal. Los resultados de este ataque indican una reducción tanto en el número de spikes como en la dispersión temporal por igual. Después, emplearon FLO sobre la misma topología y la misma configuración de neuronas, dando como resultado el incremento del número de spikes y la dispersión temporal. La investigación concluyó que ambos ciberataques generan un gran impacto, aunque JAM afecta en mayor medida el número de spikes en comparación con FLO.

Finalmente, López Bernal et al. [7] definieron una taxonomía de ocho ciberataques neuronales, incluyendo los tres mencionados previamente: Neuronal Flooding (FLO), Neuronal Jamming (JAM), Neuronal Scanning (SCA), Neuronal Selective Forwarding (FOR), Neuronal Spoofing (SPO), Neuronal Sybil (SYB), Neuronal Sinkhole (SIN), and Neuronal Nonce (NON). Estos ataques se diferencian según su comportamiento, su duración, su complejidad y el número de neuronas objetivo. Al igual que en los anteriores trabajos, han sido validados en la topología de 276 neuronas y, también, se ha utilizado el número de spikes para evaluar sus impactos. El artículo evalúa el impacto de los ataques en el corto (primeros cinco segundos) y largo plazo (últimos cinco segundos) de la simulación, indicando que JAM y NON tienen un gran impacto sobre la actividad neuronal en los primeros cinco segundos de los 27 segundos que dura la simulación, mientras que SCA y NON son más efectivos en los últimos cinco segundos.

En los últimos años, han surgido en la literatura reconstrucciones neuronales realistas, superando una de las limitaciones de las publicaciones previas. Por ello, López Madejska et al. [8] utilizaron una topología neuronal de 450 neuronas para implementar los ciberataques FLO y JAM. Dicho modelo es una simplificación del publicado por Arkhipov et al. [14], el cual representa una pequeña parte de la corteza visual primaria (V1) del ratón, concretamente la capa cuatro. Además, seleccionaron uno de los dos niveles de granularidad ofrecidos por los creadores de la topología, llamado PointNet, el cual emplea una versión simplificada del comportamiento neuronal. Los autores utilizaron la herramienta BMTK [11] para construir y gestionar el modelo, y NEST [12] para simularlo. Además, utilizan valores estáticos que representan la entrada de información desde LGN a V1, proporcionado por BMTK. En este trabajo se llevaron a cabo dos experimentos para FLO: el primero analiza cómo atacar con diferentes voltajes afecta a un número de neuronas fijado, mientras que el segundo estudia el impacto de atacar diferentes números de neuronas simultáneamente utilizando un voltaje determinado. Ambos experimentos se realizan en un instante de tiempo concreto. Por lo contrario, para JAM, se evalúa el efecto causado sobre diferentes conjuntos de neuronas en una ventana temporal. Asimismo, se utiliza la métrica del número de spikes para evaluar el impacto de ambos ataques. Los experimentos de FLO demuestran que forzar el voltaje umbral, el cual garantiza que una neurona atacada genere un spike, es la alternativa más perjudicial, y cuánto mayor sea el número de neuronas objetivo, mayor será el número de spikes. En caso de JAM, se concluye que cuanto mayor sea el número de neuronas atacadas, mayor será el impacto. Aunque este trabajo supuso un avance para comprender el impacto de ciberataques neuronales en topologías neuronales realistas, la literatura sigue presentando limitaciones en relación a la topología usada, el reducido realismo del estímulo visual empleado y la falta de investigación sobre el impacto de los ciberataques en funciones cognitivas como la visión.

Por tanto, López Madejska et al. [9] implementaron FLO y JAM en una reconstrucción completa de la corteza visual primaria del ratón, formada por alrededor de 230,000 neuronas, distribuidas en las seis capas de V1 [15]. Además, los autores utilizan tres estímulos visuales realistas, siendo estos un estímulo flash, un fragmento de película y drifting gratings con una orientación de 90 grados. Al igual que en el trabajo anterior, utilizan las herramientas BMTK y NEST para construir y simular la topología, así como el mismo nivel de granualidad, PointNet. Para evaluar el impacto causado por los ataques utilizaron el número de spikes y el porcentaje de shifts por intervalo de tiempo. Los experimentos se han enfocado en evaluar los ciberataques sobre los tres instantes más significativos de cada estímulo y han estudiado el impacto de atacar dos tamaños de neuronas sobre el total disponible (50 % y 25 % del total de las neuronas). Los resultados muestran que el impacto de FLO esta determinado por el estímulo y el instante de ataque, mientras que JAM interrumpe la actividad neuronal independiente del estímulo presentado. Por otro lado, JAM tiene mayor propagación y produce mayor porcentaje de shifts que FLO y no es necesario atacar un gran número de neuronas para producir un impacto notorio. Sin embargo, en FLO, cuanto mayor sea el número de neuronas objetivo, mayor será el impacto.

III. Diseño de la solución propuesta

Esta sección presenta la topología neuronal utilizada, el estímulo visual usado como entrada del modelo y los pasos seguidos para analizar el impacto del ciberataque seleccionado

sobre determinadas neuronas. Además, se detalla la configuración de JAM y la explicación de los parámetros de ataque usados.

III-A. Modelo neuronal y análisis del ataque

Este estudio se ha desarrollado en un modelo neuronal propuesto por Billeh et al. [15], que representa una reconstrucción detallada de la corteza visual primaria del ratón, formado por aproximadamente 230,000 neuronas con diferentes comportamientos y distribuidas en distintas poblaciones neuronales. Esta topología fue seleccionada por su alta resolución y tamaño, al incorporar un mayor número de neuronas y un nivel de detalle significativamente superior al de otros modelos existentes, tales como [14], [16], [17]. Por otro lado, los autores del modelo ofrecen dos niveles de granularidad, al igual que el modelo utilizado por López Madejska et al. [8]. El primer tipo ofrece una topología detallada a nivel biofisiológico (BioNet), mientras que el segundo es una simplificación del comportamiento de las neuronas basadas en neuronas Gated Leaky Integrate and Fire (GLIF) [18] con un enfoque point-neuron (PointNet). Para construir ambos modelos es necesaria la herramienta BMTK [11]. Por otra parte, se utiliza NEURON [19] para simular BioNet, mientras que NEST [12] se usa para PointNet. Para este trabajo, se ha escogido el modelo PointNet junto a NEST debido a que requiere menos recursos computacionales y ofrece una aproximación suficiente buena. Asimismo, este modelo simula el LGN, componente fundamental encargado de procesar el estímulo externo desde la retina y remitirlo a V1.

Si bien este componente trasmite información a la mayoría de las capas de V1, la capa cuatro (L4) es la entrada principal para el procesamiento visual inicial. Como el objetivo de esta investigación es atacar neuronas previamente identificadas para simular un estado de ceguera, se ha seleccionado el ciberataque JAM, diseñado para inhibir la actividad neuronal, con el fin de comprobar si dicho ataque genera un patrón similar a la ceguera. Además, su implementación y ejecución presentan una baja complejidad. Esta elección viene respaldada por el conocimiento aportado por un equipo de neurocientíficos pertenecientes al Instituto Universitario de Investigación en Envejecimiento de la Universidad de Murcia, quienes indicaron que una total ausencia de entrada desde LGN a L4 produce una condición similar a la ceguera. Por ello, los experimentos realizados se han enfocado en L4, específicamente en las neuronas excitadoras debido a que su abundancia permite un mejor ajuste del impacto del ataque. Aunque la estimulación de neuronas inhibidoras causaría un efecto de inhibición sobre las neuronas con las que establecen sinapsis, se prevé que su efecto global sea menor debido a su menor número. Por esta razón, el estudio se enfoca en las neuronas excitadoras, donde el ataque puede generar un impacto más significativo en la actividad neuronal. Por otra parte, para simular una ceguera en esta topología (caso base), se ha suprimido la entrada de estímulos desde LGN, dejando como única entrada a V1 la actividad externa que representa otras regiones del cerebro denominado background noise (BKG).

La Figura 1 representa la actividad neuronal sin ninguna entrada procedente de LGN, junto con el comportamiento espontáneo sin realizar ningún tipo de perturbación a las neuronas. La actividad de cada neurona, identificada por un ID y representada en el eje Y, se visualiza como un punto, cuyo color indica la población a la que pertenece. Además, los spikes están distribuidos por las capas de V1, desde L1 hasta L6. El eje X muestra el tiempo de la simulación en milisegundos y la duración de la película. Como se puede observar, el estímulo visual genera una gran actividad neuronal en el comportamiento espontáneo mientras que, en el estado de ceguera, la única actividad que aparece es la inducida por el propio BKG. En este trabajo se ha utilizado una película de 3,000 ms como estímulo visual, la cual sirve como entrada tanto al comportamiento espontáneo como a la actividad obtenida al atacar neuronas. En concreto, dicho estímulo está formado inicialmente por un estímulo neutro de 500 ms, seguido por 2.500 ms correspondientes al clip de 120 segundos de la película Touch of Evil (Zugsmith & Welles, 1958). Por otro lado, la duración de la simulación está fijada a tres segundos con una resolución de 0,25 ms, siendo suficiente para evaluar el impacto del ciberataque y su propagación.

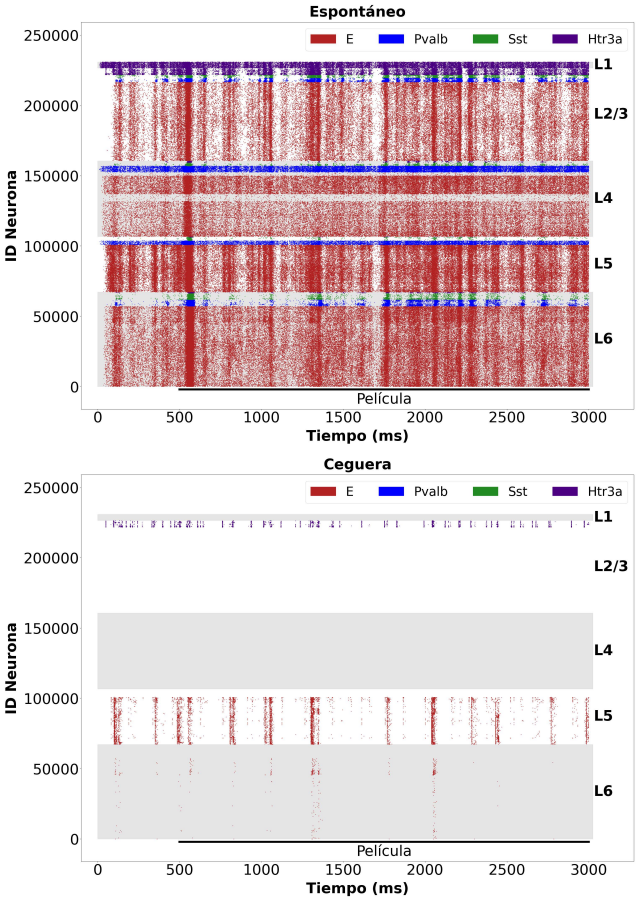


Figura 1: Comparación de la actividad neuronal entre el comportamiento espontáneo y la ausencia de estímulo visual externo simulando un estado de ceguera (caso base).

La Figura 2 muestra una representación visual de la topología utilizada, destacando el LGN y todas las capas de V1. Esta figura indica qué capa recibe la mayor cantidad de información del LGN (flecha verde continua) y, a su vez, se señalan, con un círculo rojo, aquellas neuronas seleccionadas

dentro de dicha capa para ejecutar el ciberataque JAM. Por otro lado, la figura incluye el estímulo visual utilizado como entrada a la topología, que en este caso corresponde a la película de tres segundos descrita con anterioridad. Los resultados obtenidos se guardan en formato CSV para facilitar su análisis, utilizando como métrica el número de spikes, que cuantifica el número de activaciones de todas las neuronas.

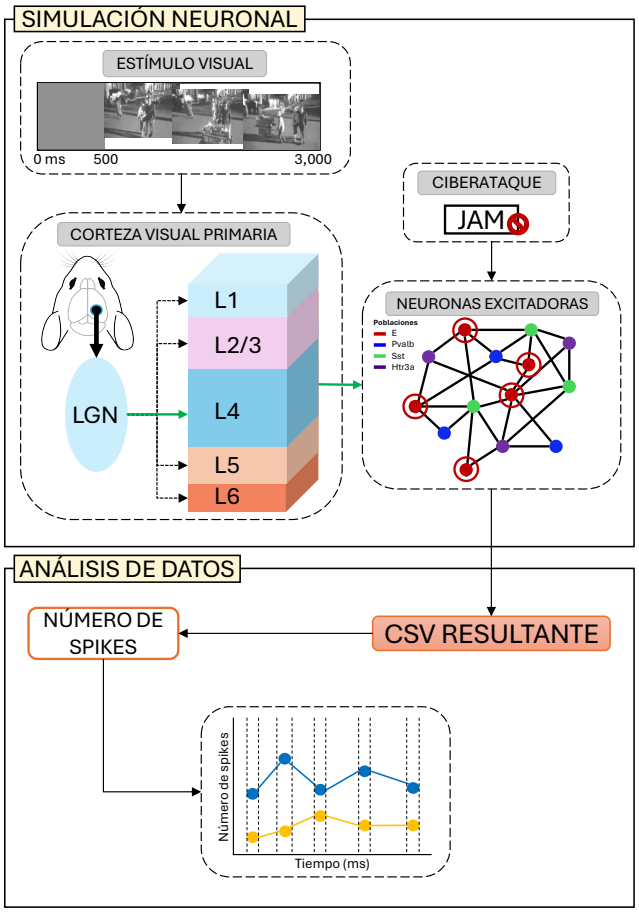


Figura 2: Representación gráfica de la topología utilizada junto a las neuronas seleccionadas, el estímulo visual externo, el ciberataque implementado y el análisis de los datos resultantes.

III-B. Configuración del ciberataque

Debido a la imposibilidad de atacar todas las neuronas excitadoras desde el punto de vista del atacante por la complejidad del propio cerebro, se seleccionaron aleatoriamente distintos subconjuntos, representando el 5 %, 10 % y 25 % del total de neuronas excitadoras (45,761 neuronas) de L4. Para estudiar la variabilidad de la topología, cada experimento se ejecutó diez veces con el objetivo de examinar si existe una diferencia notable entre ejecuciones. El número de ejecuciones se eligió teniendo en cuenta los criterios de estudios previos [6], [8]. Como se mencionó previamente, se utilizó como entrada de la topología el estímulo visual de una película, concretamente el trial número 9 de LGN y el trial 89 de BKG, en base al razonamiento expuesto por López Madejska et al. [9]. El artículo destaca que por cada estímulo visual, los autores de la topología ofrecen diez ensayos diferentes, a los cuales les corresponde un trial determinado del BKG. En cuanto a los instantes de ataque, JAM se aplica durante un intervalo de 400 ms, dado que la literatura no ha explorado ventanas temporales superiores a 250 ms. Se han seleccionado diferentes momentos en función de la variación en el número de spikes con respecto a la actividad espontánea, considerando también el intervalo en el que ocurre la transición del estímulo neutro al de la película. Por tanto, se han definido cuatro ventanas de ataque, cada una analizada de forma independiente, de tal forma que solo hay un único ataque activo por simulación:

- **500-900 ms**, donde se produce el punto crítico del cambio de estímulos de un estado visual neutro al inicio de la película.
- **1,100-1,500 ms**, siendo 1,100 ms el instante con menor número de spikes de toda la simulación.
- **1,300-1,700 ms**, el cual registra el pico de spikes más alto después del instante crítico comprendido entre 500-900 ms.
- **1,600-2,000 ms**, donde se presenta una reducción de spikes tras el instante 1,100 ms, formando un pequeño valle con el intervalo anterior.

Por último, se establece el voltaje de la membrana de las neuronas objetivo en el valor mínimo (V_reset) durante el ciberataque para evitar que generen actividad. Dicho parámetro fue escogido tras los experimentos realizados en la literatura [5], [7], donde se demuestra que los valores cercanos al voltaje de mínimo causan un efector mayor a la hora de reducir la actividad neuronal. La Tabla I muestra, de forma resumida, los parámetros del ciberataque explicados previamente, utilizados en todos los experimentos realizados.

Tabla I: Configuración de Neuronal Jamming para cada experimento.

Estímulo visual	Ventanas bajo ataque	Número de neuronas atacadas	Número de ejecuciones	Voltaje de ataque
Película: LGN trial 9, BKG trial 89	500-900 ms, 1,100-1,500 ms, 1,300-1,700 ms, 1,600-2,000 ms	5 %, 10 %, 25 % de las excitadoras de L4	10	V_reeset

IV. Ciberataque neuronal sobre la cuarta capa de V1

Esta sección analiza el impacto causado en las neuronas objetivo de la capa cuatro de V1 basado en la métrica del número de spikes de forma análoga a la literatura previa [6], [10], [8].

El formato de las gráficas que se presentan a lo largo de esta sección tiene el eje X simplificado a 30 intervalos de 100 ms cada uno para mejorar su visualización. Cada configuración se ha ejecutado de forma independiente, representándose en la gráfica con distintos colores: la línea azul corresponde al experimento en el que JAM se aplica sobre el 5 % del total de neuronas excitadoras de L4; la línea amarilla, al 10 %; y la línea verde, al 25 %. Además, estos resultados se comparan con la actividad espontánea (línea naranja) y con la representación de la ceguera (línea rosa). Tal como se indicó en la Sección III, todos los experimentos se analizan utilizando la métrica del número de spikes.

La Figura 3 muestra el impacto de JAM desde el intervalo donde se realiza el cambio del estímulo neutro al de la película (500-900 ms). Al ser el pico más alto del número de spikes de forma natural, el ataque logra inhibir el 45.90 %, 58.46 % y 67.91 % de dicha actividad espontánea, dependiendo del número de neuronas afectadas. Aun así, estos niveles de inhibición se mantienen alejados del estado de ceguera. Sin embargo, los siguientes intervalos se aproximan significativamente, sobre todo el intervalo 9. Una vez finalizados los 400 ms de ataque, el impacto se propaga hasta siete intervalos más (700 ms). Si bien las neuronas vuelven a su comportamiento espontáneo, se observa en los intervalos finales (intervalos 24 y 25) de la simulación una ligera desviación, lo que podría sugerir efectos residuales del ataque. Respecto al número de neuronas objetivo, atacar el 5 % de las neuronas genera un impacto significativo. Además, la diferencia en el efecto del ciberataque entre atacar el 5 %, 10 % o 25 % de las neuronas no es tan notable como la diferencia entre el caso espontáneo y el 5 %, demostrando que una cantidad reducida de neuronas atacadas es suficiente para generar un gran impacto. Por tanto, el número de neuronas a atacar dependerá de los objetivos del atacante en cada momento.

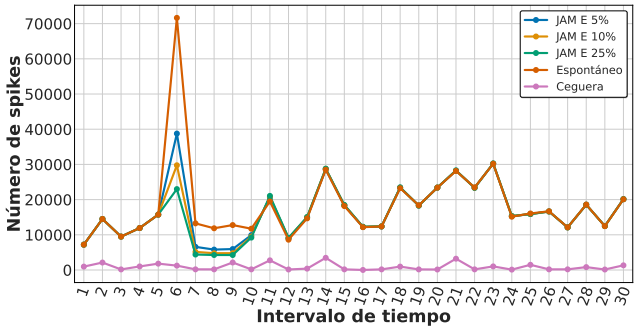


Figura 3: Número de spikes tras ejecutar JAM entre 500-900 ms en diferentes números de neuronas atacadas.

En cuanto a la Figura 4, se observa que en el intervalo 12 se produce una disminución sustancial en la actividad neuronal, con reducciones del 49.49 %, 59.22 % y 66.05 % en el número de spikes en comparación con la condición espontánea. Sin embargo, esta inhibición no se mantiene en los intervalos posteriores (13-16). Estos resultados vienen condicionados por el caso espontáneo, ya que cuanto más alta sea la actividad neuronal en la condición espontánea, mayor será la dificultad para inducir una inhibición significativa. Después de la ventana de ataque, las neuronas tardan en volver a su actividad normal alrededor de diez intervalos (1,000 ms). Al igual que en el experimento anterior, la diferencia entre los distintos conjuntos de neuronas objetivo (5 %, 10 % y 25 %) es escasa.

La Figura 5 representa la ejecución de JAM en el segundo intervalo con la mayor subida de spikes, concretamente el intervalo 14. De forma similar al primer experimento, el ataque provoca una reducción del 36.33 %, 45.18 % y 50.78 % en la actividad neuronal espontánea, en función del conjunto de neuronas objetivo. A pesar de ello, esta inhibición no alcanza completamente el estado de ceguera durante dicho intervalo. En cambio, los intervalos de ataque posteriores

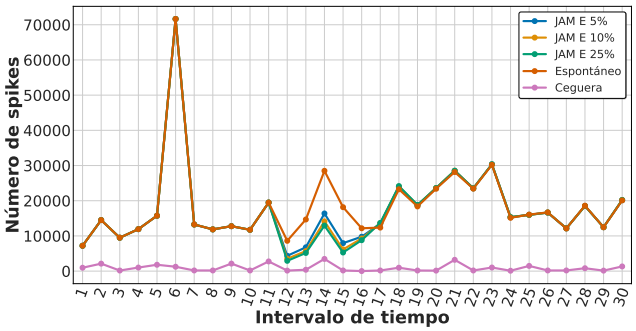


Figura 4: Impacto de JAM en el número de spikes entre 1,100-1,500 en diferentes números de neuronas atacadas.

muestran una reducción más pronunciada, siendo el intervalo 16 el más próximo a la ceguera. Esto se debe a la tendencia decreciente que tiene el caso espontáneo a partir del intervalo 15. Posteriormente, el efecto del ciberataque se propaga durante siete intervalos adicionales (700 ms) en la simulación. Respecto al impacto generado entre las diferentes neuronas objetivo, se observa que en situaciones en las que el número de spikes no es muy alto, la mejor configuración en términos de coste-beneficio es atacar el 5 %. Por el contrario, en casos donde el número de spikes es muy alto, la mejor opción es atacar números más altos de neuronas, en este caso, el 25 %.

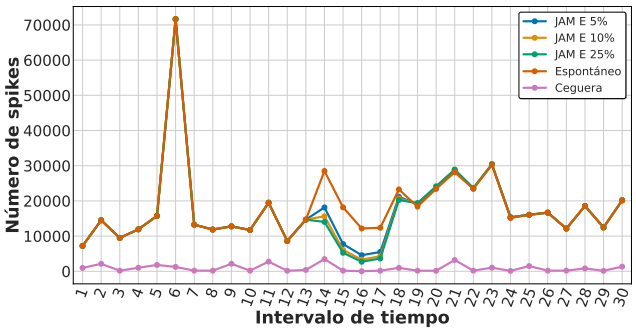


Figura 5: Variación en el número de spikes entre 1,300-1,700 ms en diferentes números de neuronas afectadas por JAM.

Por último, la Figura 6 muestra el efecto del ciberataque en un intervalo donde el número de spikes en la actividad espontánea presenta un valle. Como se observa durante la ventana de ataque, se produce una reducción del 41.74 %, 51.99 % y 59.61 % en la actividad neuronal espontánea, según las neuronas atacadas. No obstante, la distribución de número de spikes se mantiene homogénea. Aunque la reducción es significativa, es moderada en comparación con los experimentos mostrados anteriormente. Además, las neuronas tardan aproximadamente ocho intervalos (800 ms) en recuperar su actividad espontánea. Respecto al número de neuronas atacadas, aunque el impacto se mantiene similar, se aprecia una ligera diferencia entre ellas en comparación con los otros experimentos. Esto demuestra que, en intervalos con números de spikes estables, cuanto mayor sea el número de neuronas atacadas, mayor será la inhibición de la actividad neuronal.

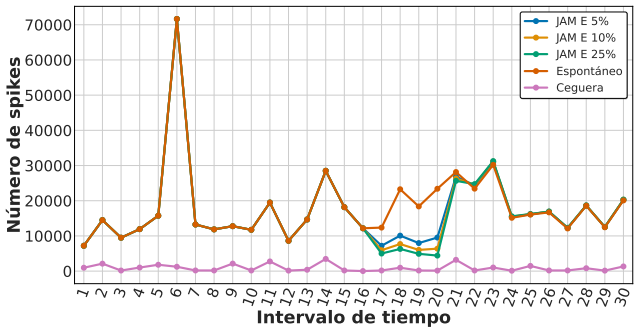


Figura 6: Efecto de JAM en la actividad neuronal entre 1,600-2,000 ms para distintos porcentajes de neuronas atacadas.

V. Discusión

En esta sección se comparan y discuten los resultados mostrados en la Sección IV, además de contrastarlos con la literatura reciente sobre ciberataques neuronales.

Respecto a la experimentación presentada en este trabajo, JAM presenta ciertas diferencias según el instante de ataque y el número de neuronas objetivo. Como se ha mencionado previamente, el objetivo es aproximar la actividad neuronal al nivel de la ceguera utilizando un ciberataque neuronal sobre un conjunto especifico de neuronas, en este caso neuronas excitadoras de la capa 4 de V1. Se ha podido observar que según la intensidad del estímulo externo sobre la actividad neuronal espontánea que se dé en el primer intervalo de ataque, se puede generar una inhibición más pronunciada, como en el intervalo 12 de la Figura 5 y el intervalo 17 en la Figura 6. En otros casos, la inhibición alcanza una reducción limitada, como sucede en el intervalo 6 de la Figura 3 y el intervalo 14 en la Figura 5. Esto sugiere que el efecto del ciberataque depende de la actividad generada en el espontáneo, siendo el ataque más efectivo cuanto más bajo es el número de spikes de la actividad neuronal basal. Por ello, el experimento que más se aproxima a la condición de ceguera es aquel en el que el ataque se ejecuta en la ventana de 500-900 ms. Aunque el ataque se inicie en un intervalo con un alto número de spikes en la actividad espontánea, los intervalos posteriores muestran una reducción en su actividad neuronal.

Respecto a la propagación, se aprecia que en los experimentos donde se ataca a los picos más altos de spikes, las neuronas tardan 7 intervalos (700 ms) en volver a su comportamiento espontáneo, mientras que en los puntos más bajos, dicha duración se alarga hasta ocho (800 ms) o diez intervalos (1,000 ms). Por otro lado, los resultados indican que atacar el 5 % de las neuronas excitadoras de L4 genera un impacto significativo sobre la actividad global de la topología, demostrando la importancia de dicha capa para el procesamiento del estímulo visual. Además, la diferencia del impacto entre atacar el 5 %, 10 % o 25 % de las neuronas es mínima en función de cuándo se produce el ataque. Si JAM se ejecuta en un intervalo con baja actividad espontánea, los resultados entre porcentajes de neuronas atacadas es similar. En cambio, cuando el número de spikes en la condición espontánea es alto, la diferenciación entre el número de neuronas objetivo es más notable. Por lo tanto, no es necesario atacar un gran número de neuronas ni a todas las neuronas de todas las capas para producir un gran impacto en la actividad neuronal global. Sin embargo, el impacto del ataque dependerá de la actividad neuronal en el estado espontáneo. Realizado el análisis de los resultados es necesario realizar una comparación con los que se han obtenido en la literatura reciente. En este estudio se emplea la misma topología descrita por López Madejska et al. [9], con aproximadamente 230,000 neuronas. Sin embargo, la literatura aborda el impacto de atacar indistintamente a toda la topología, mientras que este trabajo se centra solo en neuronas excitadoras de L4. Por otra parte, en la literatura se analizan dos ciberataques neuronales, FLO y JAM, para evaluar su impacto sobre una topología neuronal realista expuesta a diferentes estímulos visuales. En contraste, este trabajo se enfoca en un análisis detallado del ataque JAM, con el objetivo de inducir y caracterizar un posible estado de ceguera en dicha topología. Al igual que en esta investigación, la literatura examina el impacto del ataque sobre el estímulo visual de una película, además de otros estímulos. Los autores de la literatura ejecutan JAM en los intervalos 400-500 ms, 500-600 ms y 1600-1700 ms, afectando al 25 % y 50 % del total de neuronas de la topología. Los resultados muestran que el ataque genera un impacto significativo independientemente del estímulo presente en la ventana de ataque. Además, se demuestra que no es necesario atacar un gran número de neuronas, como el 50 % de la topología, para generar un impacto notorio. En este trabajo se alcanza la misma conclusión en experimentos donde el número de spikes no es muy alto. Sin embargo, los resultados indican que el impacto del ciberataque está condicionado por la actividad espontánea. Por otro lado, la propagación del ciberataque en la literatura conlleva entre seis y siete intervalos (600-700 ms), mientras que en este trabajo, el efecto se propaga como mínimo siete intervalos (700 ms) y como máximo diez intervalos (1,000 ms). Esto sugiere que, a medida que aumenta la duración de la ventana de ataque y, dependiendo de las neuronas objetivo, el impacto del ataque en la propagación se intensifica.

A pesar del avance proporcionado por este trabajo sobre la literatura, profundizando en el impacto de un ciberataque neuronal y su posible efecto en las funciones cognitivas, en este caso, la visión, esta investigación todavía presenta algunas limitaciones. En primer lugar, el estudio se centra en un único tipo de ciberataque neuronal, considerando solo la inhibición de la actividad neuronal. Además, el análisis se restringe a las neuronas excitadoras de la L4, sin evaluar el impacto en otras capas ni en las neuronas inhibidoras. Por último, la investigación examina la condición de ceguera en un solo estímulo visual realista, sin determinar si dicha condición podría generarse independientemente del estímulo presentado.

VI. Conclusiones

Este estudio analiza el impacto de JAM sobre un conjunto determinado de neuronas excitadoras de L4 para recrear un estado de ceguera en la topología. Este ciberataque se implementa sobre una reconstrucción de V1 del ratón, formado por alrededor de 230,000 neuronas con distintos comportamientos y poblaciones neuronales, distribuidas entre las seis capas de V1. Además, se utiliza un estímulo visual realista, siendo este una película. Para poder estudiar el impacto del ciberataque se configura considerando diferentes números de neuronas

atacadas, las ventanas de ataque, el número de repeticiones por cada experimento y el voltaje utilizado.

Los resultados obtenidos muestran que JAM induce parcialmente el estado de ceguera, dependiendo tanto de la actividad espontánea presente durante la ventana de ataque como del número de neuronas inhibidas. En particular, cuanto mayor sea el número de spikes del estado espontáneo, menor sera el impacto del ataque, alejándose del escenario de ceguera. Por ello, el experimento que más se aproxima a la condición de ceguera es en la ventana de ataque de 500-900 ms. Respecto a su propagación, según donde se dé la ventana de ataque, las neuronas pueden tardan en recuperar su comportamiento espontáneo ente siete y diez intervalos (700-1,000 ms). Por otro lado, en experimentos donde la actividad espontánea es baja, se observa que atacar solo al 5 % de las neuronas excitadoras de L4 genera un impacto significativo sobre la actividad global de la topología. En cambio, en experimentos con un alto número de spikes en la actividad espontánea, la opción más óptima es atacar un mayor porcentaje de neuronas. Estos resultados, también demuestran la importancia de la capa cuatro sobre el procesamiento visual.

Como trabajo futuro, se plantea identificar cuál de los ciberataques neuronales existentes en la literatura es el más óptimo para inducir un estado de ceguera de forma completa. Por otro lado, se sugiere explorar de manera comparativa diferentes duraciones o combinaciones de ataques. Además, se propone analizar qué capa de V1 y qué tipo de neuronas (excitadoras o inhibidoras) son más vulnerables ante dicha condición. De igual manera, se plantea estudiar otros estímulos visuales descritos en la literatura para determinar si el impacto de los ciberataques puede inducir la condición de ceguera, independientemente del tipo de estímulo. Finalmente, se propone estudiar qué configuración de parámetros es la más óptima para maximizar el beneficio del atacante, utilizando inteligencia artificial para ello.

Referencias

[1] E. Musk, "An integrated brain-machine interface platform with thousands of channels," *J Med Internet Res*, vol. 21, no. 10, p. e16194, Oct 2019.

[2] Q. Li, D. Ding, and M. Conti, "Brain-Computer Interface applications: Security and privacy challenges," in *2015 IEEE Conference on Communications and Network Security (CNS)*. San Francisco, CA, USA: IEEE, Sep 2015, pp. 663–666.

[3] H. Takabi, A. Bhalotiya, and M. Alohaly, "Brain computer interface (BCI) applications: Privacy threats and countermeasures," in *IEEE 2nd International Conference on Collaboration and Internet Computing*. Pittsburgh, PA, USA: IEEE, Nov 2016, pp. 102–111.

[4] L. Pycroft, S. G. Boccard, S. L. Owen, J. F. Stein, J. J. Fitzgerald, A. L. Green, and T. Z. Aziz, "Brainjacking: Implant Security Issues in Invasive Neuromodulation," *World Neurosurgery*, vol. 92, pp. 454–462, Aug 2016.

[5] S. López Bernal, A. Huertas Celdrán, G. Martínez Pérez, M. T. Barros, and S. Balasubramaniam, "Security in brain-computer interfaces: State-of-the-art, opportunities, and future challenges," *ACM Computing Surveys*, vol. 54, no. 1, Jan. 2021.

[6] S. López Bernal, A. Huertas Celdrán, L. Fernández Maimó, M. T. Barros, S. Balasubramaniam, and G. Martínez Pérez, "Cyberattacks on miniature brain implants to disrupt spontaneous neural signaling," *IEEE Access*, vol. 8, pp. 152 204–152 222, 2020.

[7] S. López Bernal, A. Huertas Celdrán, and G. Martínez Pérez, "Eight reasons to prioritize brain-computer interface cybersecurity," *Communications of the ACM*, vol. 66, no. 4, p. 68–78, mar 2023.

[8] V. M. López Madejska, S. López Bernal, G. Martínez Pérez, and A. Huertas Celdrán, "Impact of neural cyberattacks on a realistic neuronal topology from the primary visual cortex of mice," *Wireless Networks*, vol. 30, no. 1, Jan. 2024.

[9] V. M. L. Madejska, S. L. Bernal, G. M. Pérez, and A. H. Celdrán, "Neural cyberattacks applied to the vision under realistic visual stimuli," 2025. [Online]. Available: https://arxiv.org/abs/2503.08284

[10] S. López Bernal, A. Huertas Celdrán, and G. Martínez Pérez, "Neuronal jamming cyberattack over invasive bcis affecting the resolution of tasks requiring visual capabilities," *Computers & Security*, vol. 112, p. 102534, 2022.

[11] K. Dai, S. L. Gratiy, Y. N. Billeh, R. Xu, B. Cai, N. Cain, A. E. Rimehaug, A. J. Stasik, G. T. Einevoll, S. Mihalas, C. Koch, and A. Arkhipov, "Brain modeling toolkit: An open source software suite for multiscale modeling of brain circuits," *PLOS Computational Biology*, vol. 16, no. 11, pp. 1–23, 11 2020.

[12] M.-O. Gewaltig and M. Diesmann, "Nest (neural simulation tool)," *Scholarpedia*, vol. 2, no. 4, p. 1430, 2007.

[13] I. Kuzovkin, R. Vicente, M. Petton, J.-P. Lachaux, M. Baciu, P. Kahane, S. Rheims, J. R. Vidal, and J. Aru, "Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex," *Communications Biology*, vol. 1, no. 1, p. 107, Aug 2018.

[14] A. Arkhipov, N. W. Gouwens, Y. N. Billeh, S. Gratiy, R. Iyer, Z. Wei, Z. Xu, R. Abbasi-Asl, J. Berg, M. Buice, N. Cain, N. da Costa, S. de Vries, D. Denman, S. Durand, D. Feng, T. Jarsky, J. Lecoq, B. Lee, L. Li, S. Mihalas, G. K. Ocker, S. R. Olsen, R. C. Reid, G. Soler-Llavina, S. A. Sorensen, Q. Wang, J. Waters, M. Scanziani, and C. Koch, "Visual physiology of the layer 4 cortical circuit in silico," *PLOS Computational Biology*, vol. 14, no. 11, pp. 1–47, 11 2018.

[15] Y. N. Billeh, B. Cai, S. L. Gratiy, K. Dai, R. Iyer, N. W. Gouwens, R. Abbasi-Asl, X. Jia, J. H. Siegle, S. R. Olsen, C. Koch, S. Mihalas, and A. Arkhipov, "Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex," *Neuron*, vol. 106, pp. 388–403.e18, 5 2020.

[16] M. Schmidt, R. Bakker, K. Shen, G. Bezgin, M. Diesmann, and S. J. van Albada, "A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas," *PLOS Computational Biology*, vol. 14, no. 10, pp. 1–38, 10 2018.

[17] M. J. Bezaire, I. Raikov, K. Burk, D. Vyas, and I. Soltesz, "Interneuronal mechanisms of hippocampal theta oscillations in a full-scale model of the rodent ca1 circuit," *eLife*, vol. 5, p. e18566, dec 2016.

[18] C. Teeter, R. Iyer, V. Menon, N. Gouwens, D. Feng, J. Berg, A. Szafer, N. Cain, H. Zeng, M. Hawrylycz, C. Koch, and S. Mihalas, "Generalized leaky integrate-and-fire models classify multiple neuron types," *Nature Communications*, vol. 9, no. 1, p. 709, Feb 2018.

[19] M. L. Hines and N. T. Carnevale, "The neuron simulation environment," *Neural Computation*, vol. 9, no. 6, pp. 1179–1209, 1997.

Sesión Formación II

Contenido Interactivo y Accesible para el Aprendizaje de Ciberseguridad

Llanos Tobarra
Dpto. Sistemas de Comunicación y Control
ETSI Informática
UNED
Madrid, Spain
llanos@scc.uned.es

Antonio Robles-Gómez
Dpto. Sistemas de Comunicación y Control
ETSI Informática
UNED
Madrid, Spain
arobles@scc.uned.es

Rafael Pastor-Vargas
Dpto. Sistemas de Comunicación y Control
ETSI Informática
UNED
Madrid, Spain
rpastor@scc.uned.es

Roberto Hernández
Dpto. Sistemas de Comunicación y Control
ETSI Informática
UNED
Madrid, Spain
roberto@scc.uned.es

***Resumen*—El objetivo principal de este trabajo es innovar en la creación de contenido educativo altamente interactivo, accesible y reutilizable mediante el uso del proyecto H5P, aplicado al ámbito de la ciberseguridad. Además, se busca su integración en asignaturas de grado y su difusión a través de Internet. Para ello, se han desarrollado un conjunto de contenidos multimedia interactivos y accesibles basados en la tecnología H5P para la asignatura de "Seguridad", con el fin de mejorar el rendimiento académico de los estudiantes. Asimismo, se han obtenido resultados preliminares para evaluar la satisfacción y aceptación de los estudiantes con respecto a estos nuevos materiales interactivos y accesibles.**

***Index Terms*—Ciberseguridad, Accesibilidad, Reutilización de contenidos, H5P**

Tipo de contribución: *Formación e innovación educativa (límite 8 páginas)*

I. Introducción

El aprendizaje en el área de ciberseguridad no debe abordarse únicamente desde una perspectiva teórica. Es fundamental incluir ejercicios prácticos y simulaciones que permitan la aplicación del conocimiento adquirido. En nuestra institución, a través de las distintas asignaturas relacionadas con la ciberseguridad, se ofrecen múltiples recursos para el aprendizaje, así como el desarrollo de laboratorios y simuladores para los estudiantes.

Hoy en día, no es sorprendente afirmar que la ciberseguridad se ha convertido en un desafío fundamental. La creciente demanda de profesionales en esta área ha aumentado a un ritmo mucho más acelerado que la formación de especialistas calificados [1]. Nuestros futuros ingenieros deben estar preparados para hacer frente a las amenazas tecnológicas en Internet, no solo desde un enfoque teórico, sino también con experiencia práctica. Para abordar este desafío, en la UNED se ha implementado en los últimos años una línea de investigación centrada en el uso de laboratorios virtuales y remotos, un campo en el que nuestro grupo cuenta con una amplia experiencia. Como resultado, se ha evidenciado un desarrollo significativo en las habilidades de pensamiento crítico [2], [3], [4]. Además, la gamificación del proceso de aprendizaje mediante estudios de caso también ha demostrado mejoras en los resultados académicos [5], [6], [7].

El grupo de innovación CiberGID [8] lleva varios años trabajando en el desarrollo de laboratorios remotos orientados a la ingeniería. Como resultado de estos esfuerzos, se ha creado la plataforma LoT@UNED [9]. Ya en la convocatoria 2019-2020, y posteriormente en la de 2021-2022, se inició esta línea con el proyecto CiberScratch, enfocado en el aprendizaje gamificado [10]. Ahora, aprovechando el próximo cambio de plataforma educativa en la UNED, estamos considerando la necesidad de mejorar aún más nuestros contenidos, incorporando escenarios ramificados que faciliten la resolución de ejercicios guiados por los estudiantes. Asimismo, se está trabajando en la creación de nuevos materiales educativos alineados con la nueva metodología de enseñanza.

Como parte de esta mejora, buscamos evolucionar hacia un contenido educativo altamente interactivo y gamificado, basado en el proyecto H5P [11]. Esto incluirá plantillas de escenarios ramificados para la resolución guiada de estudios de caso. En este trabajo, hemos desarrollado un conjunto de recursos y escenarios con H5P para la asignatura de "Ciberseguridad", con el objetivo de mejorar el rendimiento académico de los estudiantes y fomentar su interacción durante el proceso de aprendizaje, entre otras ventajas. Los contenidos creados con esta tecnología pueden integrarse en diversas plataformas de aprendizaje como Moodle, WordPress, Canvas, Lumi Cloud, entre otras, o bien ser utilizados como recursos independientes y accesibles, como se muestra en [12]. H5P es, en esencia, un formato de contenido que permite la creación y distribución de recursos interactivos de manera sencilla y gratuita.

Además, se ha llevado a cabo una evaluación preliminar para medir la satisfacción y aceptación de los estudiantes con respecto al acceso a los recursos H5P. Se han analizado factores como la influencia social, la actitud hacia estos nuevos recursos, su facilidad de uso y utilidad percibida, así como la intención de utilizarlos en otros contextos. La accesibilidad y la interactividad de los recursos propuestos también han sido consideradas en este estudio.

La estructura del presente trabajo es la siguiente: en la Sección II se describe la metodología utilizada y las preguntas de investigación. En la Sección III se presentan los recursos interactivos y accesibles de H5P en el contexto de este estudio. La Sección IV detalla el proceso de recopilación y análisis de datos. Finalmente, en la Sección V se exponen las conclusiones y posibles líneas de trabajo futuras.

II. Metodología

Una visión general de la metodología seguida se presenta en la Figura 1. Esta metodología se estructura en cuatro fases principales:

1. *Análisis del Contexto Educativo*. En primer lugar, es fundamental considerar el contexto educativo en el diseño de cada actividad. Se analizan los resultados de aprendizaje del curso para identificar sus objetivos principales. Con base en estos objetivos, se desarrollan los recursos pedagógicos digitales.
2. *Diseño de Contenidos Educativos*. A partir del contexto educativo, se procede a describir y diseñar los recursos de aprendizaje. Luego, estos recursos se desarrollan conforme al diseño planteado y se prueban en entornos educativos.
3. *Recopilación de Datos*. En esta fase, se recolectan los indicadores seleccionados desde una perspectiva tanto cualitativa como cuantitativa. La recopilación de datos se realiza a través de encuestas que pueden ser respondidas por cualquier participante del proceso de aprendizaje (estudiantes, docentes, etc.).
4. *Análisis de Datos.* Una vez obtenida y preparada la información, los datos se analizan desde un enfoque exploratorio y confirmatorio. Esto permite extraer conclusiones y proponer mejoras en el proceso de aprendizaje de los estudiantes.

Nuestra metodología ha sido implementada en diversas asignaturas de nivel universitario. Sin embargo, para describir en detalle la aplicación de esta metodología, en este trabajo nos enfocamos en una asignatura específica: "Seguridad". Esta materia, impartida en modalidad de educación a distancia, forma parte del Grado en Ingeniería Informática, asignatura obligatoria de tercer curso de segundo cuatrimestre, y tiene una carga de 6 créditos ECTS. Su enfoque se centra en los principios de seguridad en redes desde una perspectiva práctica. La metodología de enseñanza y aprendizaje sigue los principios establecidos por el Espacio Europeo de Educación Superior (EEES), garantizando que el programa de estudio ayude a los estudiantes a desarrollar un conjunto de competencias específicas (ver Tabla I).

El principal objetivo de esta asignatura es abordar la seguridad de las infraestructuras de información desde un enfoque práctico. La relación entre los resultados de aprendizaje y los contenidos desarrollados en la asignatura se resume en la Tabla II. Esta tabla también muestra la correlación entre las competencias, los resultados de aprendizaje y los contenidos de la materia, proporcionando la base para el desarrollo de los recursos educativos.

Debido a la metodología de enseñanza a distancia, la asignatura seleccionada debe adaptar el proceso tradicional de enseñanza y aprendizaje a un entorno virtual más dinámico, fomentando una mayor interacción dentro de la comunidad educativa. Este aspecto cobra aún más relevancia en la UNED, dado el elevado número de estudiantes por curso.

Otro factor clave en el diseño de los contenidos educativos es la diversidad de temas abordados en la asignatura y su naturaleza, lo que hace imprescindible la inclusión de actividades prácticas. Para evaluar el aprendizaje, la materia cuenta con tres actividades prácticas basadas en laboratorios virtuales en línea y un examen teórico.

III. Recursos Educativos

III-A. H5P

H5P [11] es un marco de trabajo gratuito y de código abierto diseñado para la creación de contenido interactivo. Su nombre proviene de HTML5 Package, ya que se basa en un conjunto de estándares que establecen cómo se puede crear y compartir contenido interactivo.

El contenido generado con H5P se desarrolla sobre HTML5, un estándar moderno de desarrollo web, lo que permite a los usuarios crear una amplia variedad de recursos interactivos, como cuestionarios, presentaciones, juegos y simulaciones. Además, estos contenidos pueden integrarse fácilmente en cualquier página web o sistema de gestión del aprendizaje (Learning Management System, LMS).

H5P es una herramienta poderosa que mejora la experiencia de aprendizaje al fomentar la participación de los estudiantes y optimizar los resultados educativos. Su uso es intuitivo y flexible, permitiendo su personalización para adaptarse a las necesidades de cualquier usuario.

El contenido en H5P puede crearse utilizando diversas herramientas, como el editor H5P (por ejemplo, Lumi Cloud), un complemento para WordPress y una aplicación móvil. El editor H5P es una herramienta basada en la web que facilita la creación de contenido interactivo. Un ejemplo de desarrollo con el editor Lumi se muestra en la Figura 3.

Este entorno permite generar una variedad de experiencias interactivas, entre ellas:

- *Cuestionarios*. Los cuestionarios en H5P pueden utilizarse para evaluar el conocimiento de los estudiantes. Pueden diseñarse para medir conocimientos teóricos, habilidades de resolución de problemas o capacidades de pensamiento crítico.
- *Presentaciones*. Las presentaciones interactivas en H5P permiten transmitir información de manera dinámica e innovadora. Pueden incluir elementos interactivos como cuestionarios, encuestas o actividades de arrastrar y soltar.
- *Juegos*. Los juegos educativos en H5P ayudan a motivar a los estudiantes y hacen que el aprendizaje sea más entretenido. Pueden emplearse para enseñar diversos conceptos en áreas como matemáticas, ciencias o historia.
- *Simulaciones*. Las simulaciones en H5P permiten crear experiencias de aprendizaje realistas. Son útiles para enseñar conceptos complejos o proporcionar prácticas interactivas que refuercen el aprendizaje.

Se ha desarrollado contenido educativo en H5P, que se puede clasificar en tres categorías principales:

- Contenido teórico general relacionado con la asignatura. Para este caso, se han utilizado plantillas de libros interactivos, lo que ha permitido combinar material teórico y videoclases con ejercicios guiados. Un ejemplo de libro interactivo se muestra en la Figura 2.
- Escenarios ramificados. Estos han permitido la creación de pequeños juegos diseñados para ayudar a los estudiantes a prepararse con mayor confianza antes de enfrentarse



Figura 1: Esquema de la metodología seguida en este trabajo

Competencia	Tipo
BC.1 - Capacidad para diseñar, desarrollar, seleccionar y evaluar, aplicaciones y sistemas informáticos, asegurando su fiabilidad, seguridad y calidad, conforme a los principios éticos y a la legislación y normativa vigente.	Específica (Informática)
BC.5 - Conocimiento, administración y mantenimiento de sistemas, servicios y aplicaciones informáticas.	Específica (Informática)
BTEti.7 - Capacidad de comprender, aplicar y gestionar la garantía y seguridad de los sistemas informáticos	Específica (Tecnologías de la Información).
CG.1 - Competencias de gestión y planificación: Iniciativa y motivación. Planificación y organización (establecimiento de objetivos y prioridades, secuenciación y organización del tiempo de realización, etc.). Manejo adecuado del tiempo.	General
CG.2 - Competencias cognitivas superiores: selección y manejo adecuado de conocimientos, recursos y estrategias cognitivas de nivel superior apropiados para el afrontamiento y resolución de diversos tipos de tareas/problemas con distinto nivel de complejidad y novedad: Análisis y Síntesis. Aplicación de los conocimientos a la práctica Resolución de problemas en entornos nuevos o poco conocidos. Pensamiento creativo. Razonamiento crítico. Toma de decisiones.	General
CG.5 - Competencias en el uso de las herramientas y recursos de la Sociedad del Conocimiento: Manejo de las TIC. Competencia en la búsqueda de información relevante. Competencia en la gestión y organización de la información. Competencia en la recolección de datos, el manejo de bases de datos y su presentación.	General

Tabla I: Descripción de las competencias específicas, generales y transversales de la asignatura “Seguridad”

a los ejercicios del curso. Un ejemplo de este tipo de contenido se muestra en la Figura 3.

- Cuestionarios de evaluación inicial y encuestas de opinión estudiantil. Se animó a los estudiantes a responder una encuesta de opinión tras completar sus actividades prácticas con los recursos H5P.

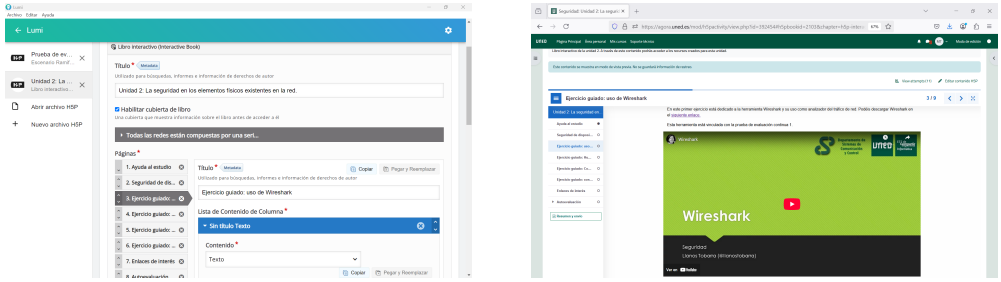
Uno de los desafíos encontrados durante este trabajo ha sido la reutilización del contenido. Aunque el formato H5P facilita en gran medida esta tarea, la actualización del material y su distribución en múltiples plataformas educativas sigue siendo un proceso complejo.

Cuando un mismo recurso se usa en diferentes cursos alojados en diversas plataformas educativas, la mejor estrategia para su integración es el uso de una plataforma en la nube. Sin embargo, este enfoque complica la recopilación de datos para el seguimiento de la actividad estudiantil.

Para abordar este problema, la combinación de H5P, xAPI y LTI ha demostrado ser una solución efectiva. Esto se logró utilizando la infraestructura de Lumi Cloud en conjunto con las herramientas de seguimiento de Moodle. Como parte de la demostración de este trabajo, algunos de estos contenidos están disponibles públicamente y de forma gratuita en un apartado específico del sitio web del Grupo de Innovación, tal y como se puede apreciar en el siguiente enlace [12].

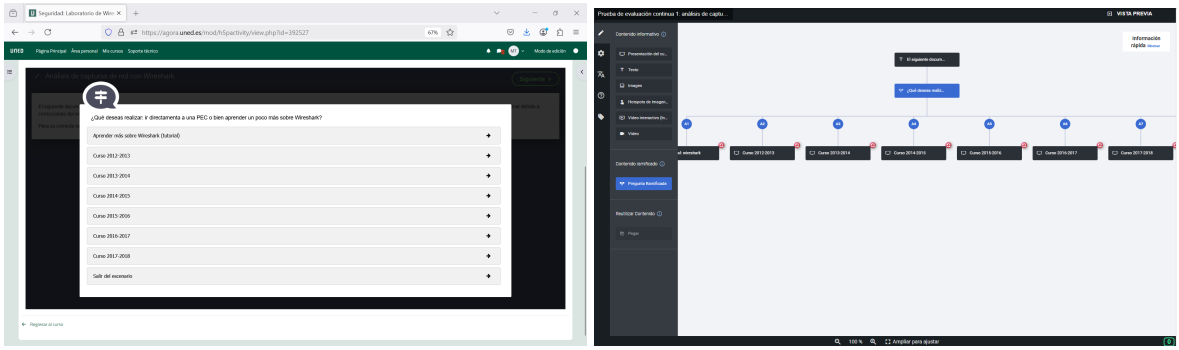
Resultados de Aprendizaje	Unidades Lectivas	Competencias
RA1. Comprender el entorno operativo de seguridad de red (NSM) y las buenas prácticas asociadas a la implementación de dicho modelo.	U1 - Principios de seguridad U2 - Seguridad Física y en Redes U3 - Seguridad Lógica U4 - Métodos de Ataque U6 - Políticas de Seguridad	BC.1 CG.1 CG.2 CG.5
RA2. Utilizar toda una gama de herramientas de software libre, entre las que se encuentran Sguil, Argus y Wireshark, para hacer prospecciones en el tráfico de red en busca de datos de contenido completo, de sesión, estadístico y de alerta.	U5 - Estrategias básicas de defensa U7 - Herramientas de Seguridad U8 - VPNs U9 y U10 - Cortafuegos U11 - Sistemas de Detección de Intrusiones (IDS) U12 - Análisis de Vulnerabilidades U13 - Diseño de Redes Seguras	BC.5 CG.1 CG.2 CG.5
RA3. Conocer y emplear las mejores herramientas para generar paquetes arbitrarios, explorar defectos, manipular el tráfico y efectuar reconocimientos.	U13 - Diseño seguro de redes U7 - Herramientas de Seguridad	BTETi.5 CG.1 CG.2 CG.5

Tabla II: Resultados de Aprendizaje (RAs), y su relación con las competencias y las unidades lectivas (U) desarrolladas en la asignatura de “Seguridad”



(a) Desarrollo de contenidos H5P (b) Contenidos desplegados en el curso virtual

Figura 2: Ejemplo de desarrollo de un recurso educativo en el editor de H5P Lumi empleando la plantilla de Libro Interactivo (a); y su posterior uso en la plataforma virtual del curso educativo de la asignatura de “Seguridad” (b)



(a) Despliegue en el curso virtual (b) Estructura del contenido del Laboratorio de Redes

Figura 3: Laboratorio de redes creado para la asignatura “Seguridad”, donde (a) muestra su apariencia dentro de la asignatura; y (b) refleja la estructura interna del escenario ramificado durante su creación

IV. Recolección de datos y análisis

Durante el curso 2023-2024, aunque se han matriculado 168 estudiantes en la asignatura, siendo un poco más del 85 % de primera matrícula, los estudiantes activos en la plataforma han sido 132 estudiantes (al menos han accedido una vez a la plataforma y han interactuado con los recursos del curso), 14 estudiantes han accedido pero no han realizado ningún tipo de actividad y 22 estudiantes no accedieron nunca a la plataforma. De los 132 estudiantes activos, sólo 117 estudiantes realizaron todas las pruebas de evaluación. Finalmente, sólo superaron la asignatura 97 estudiantes.

Para evaluar nuestra propuesta, se solicitó a los estudiantes su opinión sobre los recursos educativos desarrollados en H5P mediante una encuesta. Cada afirmación en la encuesta fue valorada utilizando una escala Likert de cinco puntos, que iba desde (5) “Totalmente de acuerdo” hasta (1) “Totalmente en desacuerdo”. Además, todas las herramientas utilizadas fueron diseñadas teniendo en cuenta los aspectos de privacidad de los datos recopilados, en cumplimiento con el Reglamento General de Protección de Datos (RGPD) de la Unión Europea [13].

La encuesta se basa en el modelo de aceptación tecnológica UTAUT [14], [15], un marco ampliamente utilizado para

analizar la adopción de nuevas tecnologías en entornos educativos. Los autores cuentan con una amplia experiencia en la aplicación de este tipo de modelos, como se detalla en [9].

Los indicadores seleccionados para el estudio fueron los siguientes:

- *Acceso a los recursos*. Evalúa si el acceso a los recursos educativos es adecuado o si presenta áreas de mejora.
- *Influencia social*. Mide cómo las opiniones de terceros afectan la percepción de los estudiantes sobre la utilidad de los recursos educativos.
- *Actitud*. Analiza la disposición de los estudiantes hacia el uso de los recursos H5P, considerando sentimientos positivos o negativos y su impacto en el rendimiento académico.
- *Facilidad de uso percibida*. Determina hasta qué punto los estudiantes consideran que el uso de los recursos H5P es sencillo y requiere menos esfuerzo en comparación con otros métodos de aprendizaje.
- *Percepción de utilidad*. Evalúa en qué medida los estudiantes creen que el uso de estos recursos mejora su desempeño en la asignatura.
- *Intención de uso futuro*. Analiza la disposición de los estudiantes a seguir utilizando los recursos H5P a lo largo del tiempo y en otras materias.
- *Interactividad*. Examina el nivel de participación e interacción que los estudiantes experimentan al utilizar los recursos H5P.
- *Accesibilidad*. Considera cómo la tecnología H5P facilita la integración de estos recursos en un entorno de educación a distancia.

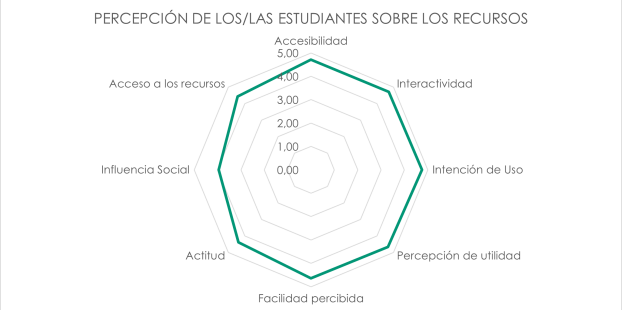


Figura 4: Percepción de los contenidos del H5P por parte de los alumnos a partir de los indicadores evaluados en la encuesta de opinión (oscila entre 5 y 1)

En cuanto a los resultados iniciales, la Figura 4 muestra los indicadores obtenidos de la encuesta de opinión para los recursos H5P propuestos a lo larga del curso 2023-2024. En este sentido, la mayoría de los indicadores medios son muy altos en general. Los mejores indicadores son Interactividad, Accesibilidad e Intención de Uso con más de 4,5 puntos sobre 5 de media. Este hecho indica que los estudiantes piensan que son muy interactivos y adecuados para otros contenidos. También valoran muy positivamente la adaptación de los materiales para su acceso a personas con diversidad funcional. Por el contrario, el indicador Actitud y el indicador de Influencia Social son inferiores al resto, aunque con buenos valores (casi 4 puntos sobre 5 de media), ya que los estudiantes no se muestran tan positivos con la nueva tecnología H5P al principio. Quizás, tienen un poco de miedo a los recursos desconocidos. Y que quizás los materiales en el curso no se valoran como un complemento al estudio, sino cómo un elemento obligatorio. Pero en general vemos que los datos son bastante buenos en general.

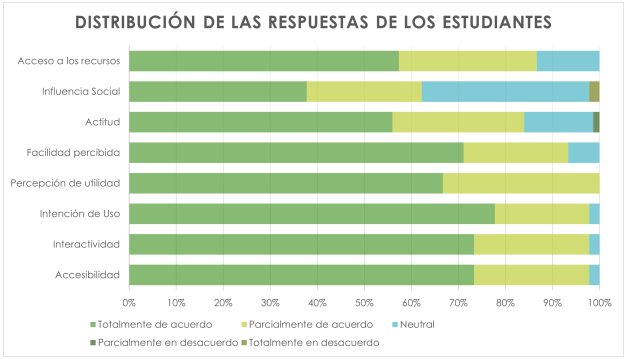


Figura 5: Distribución de respuestas de los indicadores evaluados por los alumnos en la encuesta de opinión. Oscila entre Totalmente de Acuerdo (5) y Totalmente en Desacuerdo (1)

Por otro lado, la Figura 5 muestra la distribución de respuestas de los indicadores evaluados por los alumnos dentro de la encuesta de opinión. Estos valores oscilan entre Totalmente de Acuerdo (5) y Totalmente en Desacuerdo (1). Podemos observar que son prometedores, ya que la distribución predominante es la de Acuerdo Total, así como unas pocas distribuciones son de Desacuerdo Parcial. Por esta razón, creemos que el uso de los materiales en H5P a lo largo del curso ha sido una experiencia muy positiva desde el punto de vista de los estudiantes.

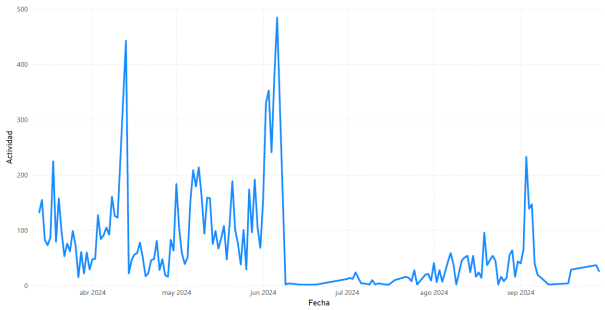
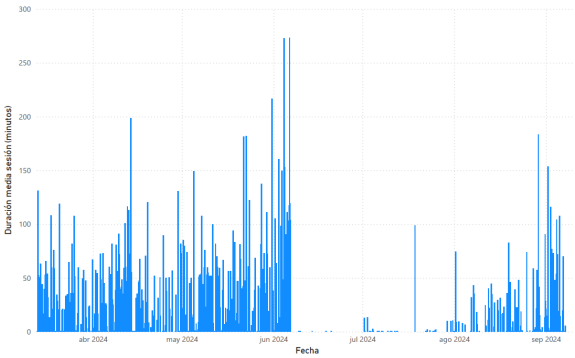
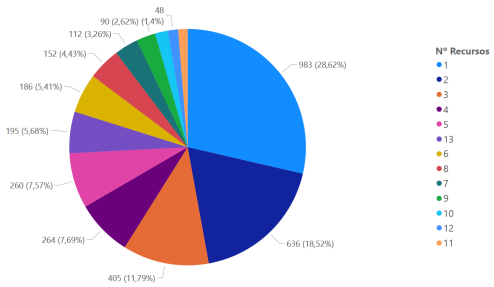


Figura 6: Evolución temporal de la actividad asociada a los recursos educativos a lo largo del segundo cuatrimestre del curso 2023-2024

Además de la encuesta de satisfacción hemos analizado los registros de actividad de acceso por parte de los estudiantes a los recursos educativos. Para ello hemos extraído el informe de actividad de la plataforma y hemos agrupado las distintas acciones en sesiones. Consideramos que se inicia una nueva sesión si han pasado 60 minutos entre una actividad y otra en un usuario. A lo largo del curso se han realizado cerca de 11928 acciones distintas de los 132 estudiantes activos. A pesar de que los estudiantes tienen a su disposición el uso de una aplicación móvil para interactuar con los cursos, sólo un 0,31 % de estas acciones han tenido lugar mediante dicha aplicación.



(a) Distribución de la duración de las sesiones a lo largo del curso



(b) Número de recursos H5P visitados por sesión

Figura 7: Gráficos orientados a la duración y el número de recursos visitados en cada sesión por parte de los estudiantes

Según podemos ver en la Figura 6, se ve claramente diversos puntos del curso académico dónde la actividad de los estudiantes es más mayor. El primero de esos picos se produce alrededor de abril, dónde los estudiantes deben realizar la primera prueba de evaluación continua. Otro pico, quizás no tan pronunciado es a principios del mes de mayo dónde se encuentra la segunda prueba de evaluación continua de la asignatura. Por último vemos un gran pico de actividad en el mes de junio coincidiendo con la evaluación ordinaria de la asignatura. Cómo cabe esperar la actividad se reduce en el mes de julio y agosto hasta prácticamente cero, para remontar levemente en el mes de septiembre, dónde tiene lugar la evaluación extraordinaria de la asignatura. Luego podemos inferir que efectivamente los estudiantes han considerado los materiales útiles de cara a preparar la evaluación de la asignatura, ya que el mayor número de accesos se concentra sobre las fechas de realización de pruebas de evaluación.

Si nos encentramos en las sesiones analizadas, cada estudiante ha tenido de media entre 12 y 13 sesiones, con una duración media de unos 15 minutos. El número de sesiones concuerda con el número de unidades asociadas a la asignatura. Con respecto a la duración de las sesiones existe gran variabilidad, conforme se puede ver en la Figura 7a, incrementándose el tiempo de las sesiones en los periodos de gran actividad ya comentados en la gráfica de actividad. En cada sesión, los estudiantes visitan entre uno y dos recursos, conforme vemos en la Figura 7b. Excepcionalmente vemos sesiones con una visita a todos los recursos del curso (13 recursos visitados por sesión). Esto puede deberse a dejar marcada la actividad dentro del curso virtual y realizar los cuestionarios de autoevaluación.

Conforme podemos ver en la Figura 8, los recursos más visitados son aquellos orientados a la parte práctica de la asignatura, los contenidos orientados a ayudar a superar los trabajos prácticas de la asignatura. De media, los recursos para los laboratorios han obtenido unas 12 visualizaciones por usuario activo en el curso, mientras que las unidades han obtenido en media entre 10 visualizaciones por usuario (unidad 9 dedicada a los Cortafuegos) hasta 5,7 visualizaciones por usuario activo (unidad 2 dedicada a la seguridad física y de redes).

Finalmente, vemos con estos datos que efectivamente los

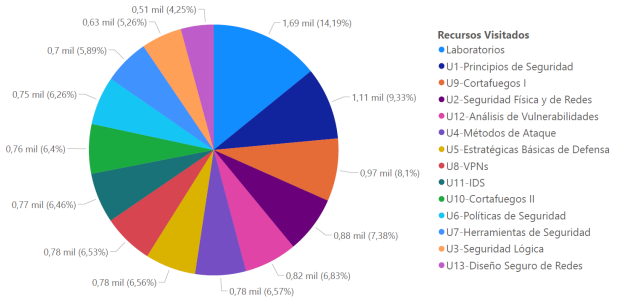


Figura 8: Distribución de las visitas a los distintos recursos H5P

estudiantes han hecho uso de los recursos, y los han encontrado útiles y fáciles de usar, especialmente de cara a la evaluación de sus conocimientos. De manera que han mostrado una actitud positiva hacia ellos, conforme reflejaban las encuestas.

V. Conclusiones

Este trabajo ha presentado la implementación de nuevos contenidos educativos altamente interactivos, accesibles y reutilizables mediante la tecnología H5P en la asignatura de "Seguridad". Además, se ha mostrado cómo estos recursos se han integrado en la materia con el objetivo de analizar la percepción de los estudiantes sobre el impacto de estos nuevos materiales interactivos y accesibles.

La incorporación de herramientas de código abierto no solo democratiza el acceso a estos recursos, sino que también garantiza que los contenidos interactivos se mantengan a la vanguardia de las prácticas educativas innovadoras. Se ha realizado una evaluación preliminar de la tecnología H5P, obteniendo resultados muy prometedores.

Como trabajo futuro, se está llevando a cabo una evaluación más profunda con el propósito de validar estadísticamente los avances científicos mediante modelos de aceptación, tanto desde un enfoque exploratorio como confirmatorio. También se planea ampliar la oferta de estos recursos y extender su uso a otras asignaturas, ya sean de temáticas similares o diferentes.

Agradecimientos

Los autores desean agradecer el apoyo del Gobierno de España (a través del Instituto Nacional de Ciberseguridad de España, INCIBE), con el Proyecto de Investigación Estratégico "Análisis de aplicaciones móviles desde la perspectiva de la protección de datos: Ciberprotección y Ciberriesgos de la información ciudadana", así como la Cátedra Internacional "Smart Rural IoT y Entornos Seguros", en el contexto del Plan de Recuperación, Transformación y Resiliencia financiado por la Unión Europea (NextGenerationEU/PRTR). Asimismo, los autores agradecen el apoyo del grupo de investigación UNED CiberCSI y del grupo de innovación UNED CiberGID.

Referencias

[1] ISACA. (2022, Jun.) Survey, R.C. State of Cybersecurity 2022. Online. [Online]. Available: https://www.isaca.org/go/state-of-cybersecurity-2022

[2] D. Dasgupta, D. M. Ferebee, and Z. Michalewicz, "Applying puzzle-based learning to cyber-security education," in *Proceedings of the 2013 on InfoSecCD '13: Information Security Curriculum Development Conference*, ser. InfoSecCD '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 20–26. [Online]. Available: https://doi.org/10.1145/2528908.2528910

[3] B. Martini and K.-K. R. Choo, "Building the next generation of cyber security professionals," *ECIS 2014 Proceedings - 22nd European Conference on Information Systems*, 01 2014.

[4] D. Willingham, "Critical thinking why is it so hard to teach?" *Arts Education Policy Review*, vol. 109, 08 2010.

[5] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? – a literature review of empirical studies on gamification," in *2014 47th Hawaii International Conference on System Sciences*, 2014, pp. 3025–3034.

[6] J. Cano, R. Hernández, S. Ros, and L. Tobarra, "A distributed laboratory architecture for game based learning in cybersecurity and critical infrastructures," in *Proceedings of 2016 13th International Conference on Remote Engineering and Virtual Instrumentation, REV 2016*. Institute of Electrical and Electronics Engineers Inc., 2016, pp. 183–185.

[7] S. Martin, G. Diaz, M. Castro, and M. Rodriguez-Artacho, "Increasing engagement in a network security management course through gamification," in *IEEE Global Engineering Education Conference, EDUCON*, vol. April-2019. IEEE Computer Society, 2019, pp. 1380–1383.

[8] CiberGID. (2024, Mar.) Grupo de Innovación Docente en Ciberseguridad. Online. [Online]. Available: https://blogs.uned.es/cibergid/

[9] A. Robles-Gómez, L. Tobarra, R. Pastor-Vargas, R. Hernández, and J. M. Haut, "Analyzing the users' acceptance of an iot cloud platform using the utaut/tam model," *IEEE Access*, vol. 9, pp. 150 004–150 020, 2021.

[10] L. Tobarra, A. Utrilla, A. Robles-Gómez, R. Pastor-Vargas, and R. Hernández, "A cloud game-based educative platform architecture: The cyberscratch project," *Applied Sciences*, vol. 11, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/2/807

[11] H5P group. (2025, Mar.) H5P. Create, share and reuse interactive HTML5 content in your browser. Online. [Online]. Available: https://h5p.org/

[12] CiberGID. (2025, Mar.) H5P Cybersecurity Resources. Online. [Online]. Available: https://blogs.uned.es/cibergid/contenidos-interactivos-para-el-aprendizaje-en-ciberseguridad/

[13] European Union. (2016) EP and the CEU: Regulation (EU) 2016/679 GDPR. Online. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[14] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003. [Online]. Available: http://www.jstor.org/stable/30036540

[15] R. J. Holden and B.-T. Karsh, "The technology acceptance model: Its past and its future in health care," *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 159–172, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046409000963

Experiencias de formación en ciberseguridad usando portales cautivos

Javier Muñoz-Calle, Francisco José Fernández-Jiménez, Rafael Estepa, Vicente Mayor, José M. Garcia-Campos
Dpto. Ingeniería Telemática, Escuela Superior de Ingenieros, Universidad de Sevilla
fjmc@us.es, fjfj@us.es, rafaestepa@us.es, vmayor@us.es, jgarcia111@us.es

Resumen- **Este trabajo describe una experiencia de formación y educación básica en ciberseguridad desarrollada en el marco de las jornadas de puertas abiertas de la Escuela Técnica Superior de Ingenieros de la Universidad de Sevilla para estudiantes de bachillerato. El objetivo es que el alumnado sea consciente de los riesgos asociados al uso de sistemas conectados, especialmente cuando se utilizan infraestructuras de acceso gratuito. Se describe el escenario experimental desplegado, la secuencia de actividades realizada, que incluye acciones de motivación y acceso aparentemente inofensivo, y los resultados obtenidos.**

***Index Terms*- Jornadas, Ciberseguridad, Suplantación.**

Tipo de contribución: *Formación e innovación educativa.*

I. Introducción

La sociedad actual está inmersa en un proceso de digitalización imparable, que obliga a los usuarios a usar tecnologías que evolucionan rápidamente, sin que estos sean realmente conscientes de su funcionamiento. El usuario medio desconoce en qué consisten las telecomunicaciones, cómo y con quién está intercambiando datos, y quiénes procesan esos datos. No se exige ninguna prueba de conocimiento para navegar por Internet o usar redes sociales, dejando al usuario la responsabilidad de aprender por sus propios medios.

Los jóvenes nativos digitales son usuarios desde sus primeros años. Las generaciones anteriores prácticamente no han recibido educación en ciberseguridad, por lo que no pueden transmitir estos conocimientos a sus hijos. Mientras que es tradición educar sobre peligros físicos (educación vial, evitar lugares peligrosos, precaución con personas desconocidas …) no lo es advertir sobre peligros cibernéticos similares. De hecho, con frecuencia los hijos poseen un mayor conocimiento del mundo digital que sus progenitores por su experiencia de uso sin restricciones, a diferencia de lo que sucede en el mundo físico. Esto les conduce a una falsa percepción de control, lo que les hace más vulnerables ante los ataques digitales. Esta situación se agrava aún más en el caso de los adolescentes, cuya mayor propensión a asumir riesgos y experimentar les lleva a afrontar situaciones más comprometidas.

Los ciberdelitos aumentan cada año [1]. Según el balance de criminalidad de 2023 del Ministerio del Interior de España [2], estos aumentaron un 25,5% respecto al año anterior y un 508,1% respecto al año 2016. Los ciberdelincuentes ven rentable y fácil cometer este tipo de crímenes. Solo precisan tener más conocimientos que el usuario (el equivalente a ser más fuertes en el mundo real, ya que aquí la fortaleza es el conocimiento), sin necesidad de enfrentarse físicamente, y viéndose amparados por el anonimato. Asímismo, tienen capacidad para realizar ataques indiscriminados a gran escala y lucrarse con el comercio de información sensible. Esta situación se ha acentuado con el auge de la Inteligencia Artificial (IA), cuyo uso ha propiciado que se perfeccionen los ciberdelitos y su detección resulte aún más compleja [3].

Las empresas tratan de proteger a sus usuarios proporcionando sistemas de seguridad cada vez mejores, pero siempre hay un compromiso entre facilidad/experiencia de uso y seguridad. El usuario también debe hacerse responsable, aprender y exigir a las empresas medidas de seguridad más eficaces. Por ejemplo, hasta hace poco el uso de HTTPS (*HyperText Transfer Protocol Secure*) no estaba extendido, debido al coste que supone obtener certificados digitales. Gracias a las presiones de los navegadores (mostrando avisos), buscadores (degradando las páginas inseguras en los resultados) y los usuarios, la mayor parte de la web usa en la actualidad HTTPS.

A pesar de ese esfuerzo, el usuario sigue suponiendo el eslabón más débil en ciberseguridad. Incluso con la mejora de la tecnología, el error humano representa la vulnerabilidad más habitual de la que se aprovechan los ciberdelincuentes [2]. Muchas personas confían en las infraestructuras de acceso y no tienen reparo en usar puntos de acceso wifi públicos u ordenadores compartidos si son gratuitos. Los usuarios confían que otras personas con más conocimientos han comprobado que son seguros y no se ven como objetivos potenciales de los ciberdelincuentes, suponiendo que su información no tiene valor para ellos. Sin embargo, ignoran que muchos ataques están automatizados o persiguen suplantar su identidad con múltiples fines maliciosos, frecuentemente económicos o sociales. El riesgo al que están sometidos los usuarios se ha agravado aún más con la implantación de la tecnología *Internet of Things* (IoT) y el incremento de dispositivos conectados a Internet que ha conllevado [4]. En [5], se muestra una taxonomía de los daños provocados por los ciberdelitos, donde se incluyen daños de pérdida de datos, económicos, psicológicos, reputacionales y sociales.

En los últimos años se observa un aumento de campañas de concienciación, tanto de entidades públicas como INCIBE [6], dirigidas al público general, como de entidades privadas para sus trabajadores o usuarios, siendo especialmente habituales las llevadas a cabo por las entidades bancarias, por el daño que sufren tanto económico como reputacional. Sin embargo, muchas veces estas campañas de concienciación llegan tarde, cuando ya se han producido los ciberdelitos, o tienen un bajo impacto por la falta de interés de los usuarios. Es necesario concienciar al usuario antes de usar la tecnología.

La Dirección de la Escuela Superior de Ingeniería de la Universidad de Sevilla, con la ayuda de sus Departamentos, organiza cada año las Jornadas de Puertas Abiertas durante las cuales estudiantes de centros de educación secundaria realizan

visitas para conocer las enseñanzas impartidas en la escuela [7]. Durante las jornadas llevadas a cabo el curso 2024-25, el Departamento de Ingeniería Telemática participó los días 4 y 6 de febrero, desarrollando en sus laboratorios una experiencia de formación, educación y concienciación en ciberseguridad, de 12 minutos de duración, en la que participaron un total de 289 estudiantes divididos en ocho sesiones. En la experiencia se mostraba la relevancia de las telecomunicaciones, la forma en que se transportan los datos por las redes y cómo se puede acceder a información confidencial fácilmente. Como fin último, se pretendía concienciar a los asistentes sobre la importancia de la ciberseguridad, conocer los riesgos a los que están expuestos los usuarios digitales y qué medidas deben adoptar para evitarlos.

Se perseguía realizar una experiencia de formación que resultase asimilable para el alumnado de bachillerato, usuarios nativos digitales pero que, en general, no poseen una adecuada formación en el campo de la ciberseguridad y las TIC. Asimismo, la naturaleza de las jornadas en las que se enmarca la experiencia exige que esta sea de una duración reducida. En la literatura podemos encontrar múltiples propuestas de experiencias de formación en ciberseguridad. [8] y [9] analizan distintos enfoques y técnicas que se han empleado, como talleres interactivos, simulaciones de ataques y plataformas de aprendizaje gamificado entre otras, analizando sus ventajas e inconvenientes. No existen propuestas genéricas, ya que cada experiencia se orienta según el perfil del usuario receptor y el tiempo y recursos disponibles. Algunas de ellas están especializadas en usuarios con más formación que los nuestros, requieren varias sesiones para su realización o muchos recursos. Por ejemplo, en [10] la formación está enfocada en trabajadores en entornos corporativos y está compuesta por 5 módulos centrados en ataques de ransomware; [11] estudia el impacto de programas de formación de ciberseguridad específicos para instituciones financieras de Irlanda; y [12] utiliza una plataforma de gamificación basada en realidad virtual para formar en ciberseguridad. En este artículo proponemos una experiencia real en primera persona que, a pesar de estar orientada a usuarios digitales con poca formación en ciberseguridad, realizarse en un breve espacio de tiempo de pocos minutos y requerir una cantidad reducida de recursos, ha logrado un alto impacto en los asistentes, incentivando su motivación en la necesidad de mejorar sus conocimientos en ciberseguridad y aportándoles algunos recursos mínimos de defensa. Frente a otras propuestas, la nuestra destaca por requerir pocos recursos, lo que facilita su despliegue.

Para lograr este propósito, se realizaron diversas actividades organizadas en dos etapas principales. En una primera se analizaron los flujos de información generados por los dispositivos de los usuarios hacia servicios habituales. En la segunda etapa se capturó y mostró información sensible de los usuarios mediante servicios "fake". Para ello, se desplegó un portal cautivo sobre una infraestructura superpuesta a la de la red del laboratorio.

Este artículo se estructura de la siguiente manera. La Sección II describe la topología y dispositivos del entorno preparado para llevar a cabo la actividad. Las etapas que componen la experiencia se detallan en la Sección III, incluyendo las demostraciones realizadas por los profesores y las acciones efectuadas por los asistentes. La Sección IV muestra los resultados experimentales obtenidos. Por último, la Sección V resume las principales conclusiones extraídas de la experiencia.

II. Despliegue del escenario

La experiencia se ha desarrollado en uno de los laboratorios del Departamento de Ingeniería Telemática. La sala está dotada con 30 ordenadores personales y un equipo con capacidad para virtualizar múltiples servidores mediante tecnología KVM (Kernel-based Virtual Machine), todos ellos interconectados con conmutadores de red. Sobre este entorno se desplegó la

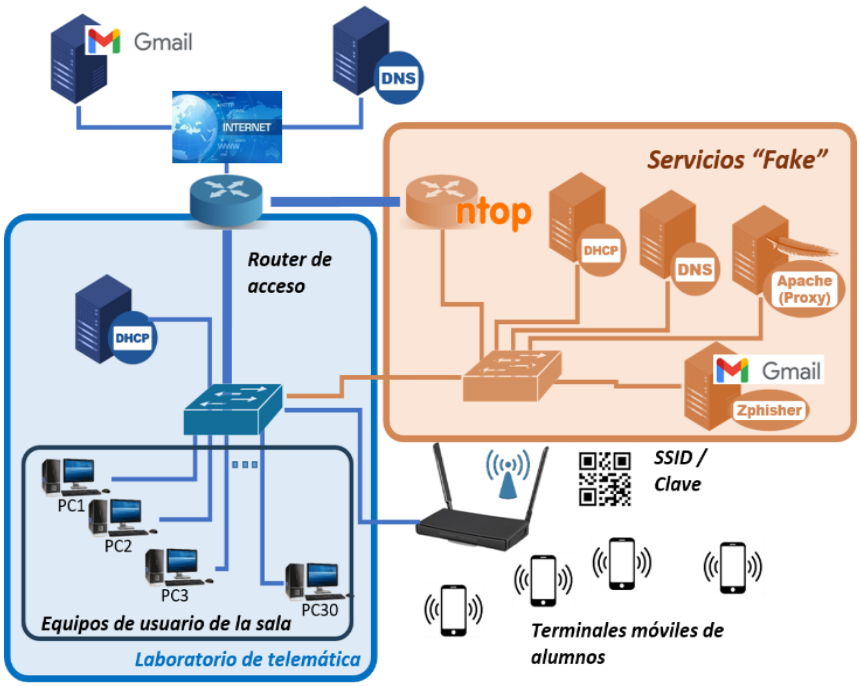


Fig. 1. Esquema de la red usada en la experiencia

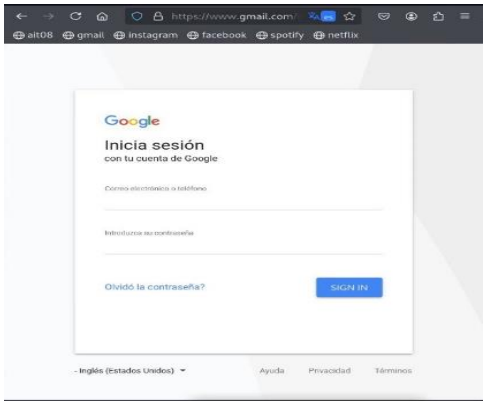


Fig. 2. Página web suplantada (portal cautivo)

infraestructura necesaria para dar soporte al experimento, cuyo esquema simplificado se muestra en la Fig. 1.

En su configuración habitual (recuadro "Laboratorio de telemática"), los *equipos de usuario de la sala* obtienen su configuración de red del *servidor DHCP (legítimo)* del laboratorio. Este servidor asigna a cada equipo sus parámetros de red (dirección IP y máscara), ajusta su tabla de encaminamiento para que tenga capacidad de acceso al exterior (Internet) a través del *router* (legítimo) de la sala, y le asigna un servidor público de resolución de nombres de dominio (*Servidor DNS legítimo*).

Para dar cobertura a la experiencia de ciberseguridad, se añadieron múltiples elementos (recuadro "Servicios Fake"), representados en la Fig. 1. En los siguientes subapartados se resume el propósito de cada uno de ellos, agrupados por la funcionalidad aportada.

Para facilitar que la experiencia pueda ser reproducida con facilidad en otros entornos, los ficheros de configuración y scripts modificados de las distintas herramientas empleadas se encuentran disponibles para la comunidad en el repositorio https://github.com/javmunca2/experiencia_phising/. Los ficheros, organizados en directorios conforme a la herramienta a que corresponden, preservan el nombre con que deben ser aplicados.

A. *Monitorización de la red y geolocalización*

En una primera etapa de la experiencia, se muestra a los usuarios la ubicación geográfica de los servicios a los que acceden cuando utilizan sus dispositivos. Ello requiere capturar el tráfico generado hacia Internet, tanto a través de los equipos de la sala, como mediante sus propios terminales móviles. Para dar soporte a este requisito, se instalaron los siguientes componentes:

- *Punto de acceso Wi-Fi*: ofrece una red Wi-Fi a la que pueden acceder los terminales móviles. Fue configurado en modo puente, de modo que los terminales trabajasen en el mismo segmento de red y fuesen configurados por el mismo servidor DHCP, que el resto de equipos de usuario de la sala.

```
Gmail Username: gdf******** Pass: ********df
Gmail Username: cca******** Pass: ********ad
Gmail Username: ssa******** Pass: ********d1
Gmail Username: rrr******** Pass: ********pa
...
```

Fig. 3. Fichero de credenciales capturadas

- *Router fake y sonda ntopng*: para monitorizar todo el tráfico de red generado por los usuarios hacia el exterior, se configuró un segundo encaminador, que suplantaba al legítimo. Asimismo, en dicho encaminador se configuró una sonda de la herramienta ntopng [13] dotada de la base de datos de geolocalización DB-IP Lite [14], encargada de analizar el tráfico saliente y de generar mapas de conexiones, accesibles vía web. A su vez, este encaminador falso reenviaba el tráfico al legítimo para que fuese cursado al exterior de forma transparente a los usuarios.
- *Servidor DHCP fake*: permite controlar la configuración de red de los equipos de usuario. En concreto, tiene dos propósitos esenciales para el experimento. Por un lado, que los equipos utilicen como pasarela el *Router fake*, permitiendo monitorizar todo el tráfico. Por otro, lograr que para la resolución de nombres utilicen el *servidor DNS fake*, en lugar del legítimo, requisito necesario para el ataque de suplantación web. Aunque en el laboratorio, al ser un entorno controlado, no fue necesario, en redes públicas o compartidas se podría lograr que los equipos de usuario utilicen el *servidor DHCP fake*, en lugar del legítimo, inutilizando este último mediante un ataque de inundación DHCP.

B. *Suplantación web mediante portal cautivo*

En su segunda etapa, el experimento persigue realizar un ataque de suplantación web. En concreto, se pretende suplantar la página web del servidor de correo electrónico Gmail [15], ampliamente usada en general, y en especial por el alumnado de bachillerato, habitualmente acostumbrado a trabajar con la plataforma educativa Google Classroom. Para dar cobertura a este objetivo, se emplearon los siguientes componentes:

- *Servidor de Phishing (Gmail fake):* ofrece una página web con un aspecto similar al de la página oficial de Gmail (Fig. 2). Se empleó la herramienta de suplantación automatizada Zphisher [16], capaz de ofrecer páginas web clonadas (portales cautivos) de múltiples servicios públicos, desplegadas sobre un servidor web con soporte de lenguaje PHP. Asimismo, se encarga de registrar las credenciales introducidas en dichas páginas web y, posteriormente, de redirigir al usuario a la web legítima del servicio correspondiente.

  Para mejorar el impacto de la experiencia y respetar la protección de datos de los usuarios, la herramienta fue ligeramente modificada. Se alteró parte de la lógica (escrita en lenguaje POSIX Shell) para que las credenciales capturadas no fueran ni impresas en pantalla ni almacenadas en disco de forma completa, sino que se almacenase una versión parcial de las mismas, en la que únicamente se recogiesen los primeros tres y últimos dos

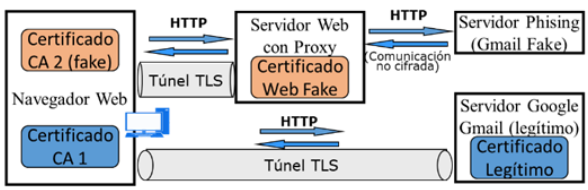


Fig. 4. Certificados empleados y conexiones.

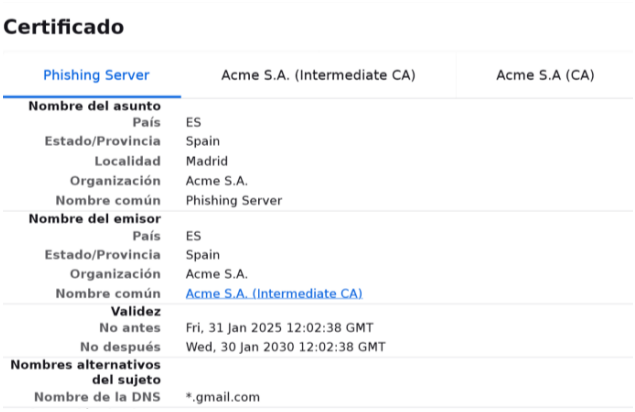


Fig. 5. Datos del certificado del Servidor de phishing

caracteres del usuario y la clave, respectivamente, permitiendo así que cada usuario pueda identificar sus propias credenciales, pero sin poner en riesgo su privacidad (Fig. 3).

- *Servidor web proxy*: la herramienta Zphisher solo soporta conexiones web no seguras (protocolo HTTP), lo que dificulta su uso en escenarios reales, dado que en la actualidad tanto navegadores como servidores rechazan este tipo de conexiones (la URL real de Gmail sólo admite conexión segura HTTPS). Los navegadores más recientes vienen configurados para tratar de conectarse por defecto siempre mediante HTTPS y muestran avisos cuando el servidor no admite conexiones seguras. Además, muchas páginas emplean la especificación HSTS (*HTTP Strict Transport Security*) para obligar al navegador a usar conexiones seguras en las siguientes conexiones. Incluso, los navegadores traen precargada una lista de páginas con HSTS por defecto. Para solventar esta limitación y lograr una experiencia realista, se configuró un servidor web Apache [17], dotado del certificado necesario para que actuase como proxy entre los navegadores de usuario y el servidor de phishing. Dicho servidor terminaría la conexión segura HTTPS (HTTP sobre TLS) con los navegadores, continuándola de manera no segura hasta el servidor de phishing.
- *Servidor DNS fake*: permite controlar la resolución del dominio a suplantar "www.gmail.com", de modo que sea traducido a la dirección IP del *servidor web proxy* en lugar de a la dirección del *servidor legítimo de Google Gmail*.

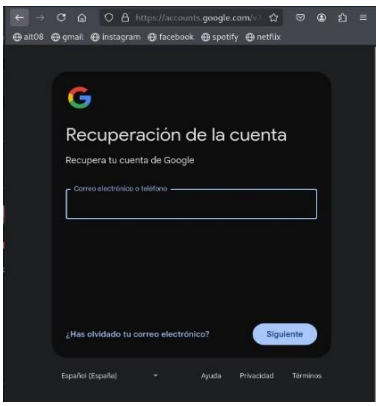


Fig. 6. Página web legítima



Fig. 7. Motor de ciberseguridad Shodan

- *Equipo de usuario con navegador comprometido:* en el navegador de los equipos de trabajo de la sala (Firefox versión 134.0.2), junto al certificado de la Autoridad Certificadora (CA) raíz del *servidor web legítimo*, se instaló el certificado de la CA raíz del *servidor web proxy*, De este modo, se logra que al acceder de forma segura al dominio "www.gmail.com", el navegador admita como válidos, sin presentar avisos de alarma, tanto el certificado presentado por el *servidor web proxy* (a través del cual se accede a la página web suplantada) como el del servidor web legítimo.

En la Fig. 4, se muestra la configuración usada de certificados digitales en navegadores y servidores, así como las conexiones HTTP y HTTPS que se establecen entre ellos. El certificado digital del servidor proxy se ha generado creando una infraestructura de clave pública (PKI, *Public Key Infrastructure*). Respetando el comportamiento de la CA de confianza que suelen traer incorporadas los navegadores, se ha creado una CA raíz exclusivamente para este experimento, cargada en el navegador de los equipos de usuario de la sala como una CA de confianza para identificar sitios web. Cualquier certificado (no expirado) firmado por esta CA raíz será tomado como válido y confiable. Con esta CA se ha firmado el certificado de una CA intermedia (usadas frecuentemente para minimizar el uso de las claves privadas de la CA raíz, reduciendo así su exposición a revelación o robo, y permitiendo revocar su certificado en caso de compromiso sin necesidad de modificar los clientes). A su vez, esta CA intermedia se ha utilizado para firmar el certificado digital instalado en el *servidor web proxy*, asociándolo al conjunto de dominios "*.gmail.com".

Todos los datos del certificado del servidor web legítimo podrían falsificarse, salvo sus claves públicas y las huellas digitales. Sin embargo, en el certificado generado se prefirió usar valores diferentes, con nombres explícitos (tales como

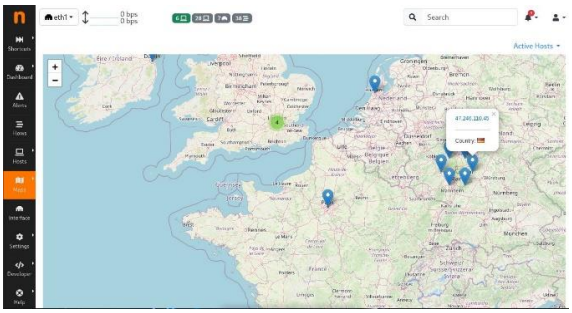


Fig.8. Sonda ntopng con tráfico de servicios en Europa



Fig. 9. Sonda ntopng con tráfico de los terminales móviles

"Servidor Phishing", Fig. 5), para que la experiencia permitiese determinar cuántos usuarios toman la precaución de verificar el certificado digital cuando acceden a una web en la que introducen datos personales.

Si se hubiese usado el navegador de los terminales móviles de los participantes, al acceder al servidor de phishing mediante la dirección "www.gmail.com", al no estar comprometidos, habría alertado de certificado no válido. Así mismo, en dispositivos actualizados, estos navegadores suelen disponer de manera precargada de la dirección de múltiples servidores (lista HSTS precargada), entre ellos del servidor Gmail, por lo que no habrían sido afectados por el servidor DNS falso. Por ambos motivos, no se usaron para la prueba del ataque de phishing, empleándose en su lugar los equipos del laboratorio.

Con la configuración expuesta, se logra que cuando los usuarios intentan acceder desde los navegadores de los equipos de la sala a la URL legítima "https://www.gmail.com/", su solicitud sea inicialmente atendida por el servidor de phishing (Fig. 2). Una vez que el usuario introduce sus credenciales y pulsa en el botón de acceso, sus datos son capturados y es redirigido al servidor oficial de correo. Éste, además de actualizar el dominio a "https://accounts.google.com/", solicita nuevamente las credenciales en un intento, aparentemente legítimo, de recuperación del acceso, lo cual es percibido por los usuarios como un comportamiento normal al estar accediendo a su cuenta desde un equipo distinto al habitual (Fig. 6).

III. Desarrollo de la actividad

En este apartado se describen las etapas en las que se organizó la experiencia. Cada estudiante, o par de estudiantes cuando lo requería la ocupación de la sala, estaba sentado con un puesto de trabajo (ordenador del laboratorio) a su disposición, además de disponer de su dispositivo móvil personal. El profesor, empleando un equipo con una configuración similar a la de cada estudiante, utilizó una pantalla de proyección para la exposición de los contenidos. La experiencia, con una duración aproximada de 12 minutos, se



Fig.10. Página de inicio de navegadores (equipos de la sala)



Fig.11. Mapa de ataques Kaspersky

organizó en las siguientes fases:

A. *Explicación del funcionamiento de Internet*

Inicialmente, se introduce al alumnado en el funcionamiento básico de Internet, se le explica cómo sus datos de usuario viajan constantemente por las redes, y se enfatiza la importancia de tomar precauciones de seguridad en el entorno digital.

B. *Supervisión de seguridad de dispositivos y servicios*

Como ejemplo práctico, se comenta la existencia de múltiples herramientas sencillas de supervisión de seguridad, presentando como caso particular el motor de búsqueda Shodan [18]. Este servicio web se encarga de rastrear los dispositivos conectados a Internet, registrando sus datos y posibles configuraciones de acceso inseguras. Se realiza una demostración rápida de uso de la herramienta, en la cual se localizan diversas cámaras de vigilancia con acceso abierto a Internet (Fig. 7), mostrando en vivo la imagen de cuatro de ellas.

C. *Geolocalización de servicios accedidos en Internet desde dispositivos móviles*

En segundo lugar, se explica al alumnado cómo los dispositivos de usuario actuales están constantemente enviando información, incluso sin la intervención del propio usuario, a servidores ubicados en los diferentes puntos del planeta.

Bajo la configuración de red de la sala, todo el tráfico generado por los dispositivos de usuario fijos, así como por los terminales móviles conectados a la red Wi-Fi suministrada, es cursado a través del router fake, siendo registrado por su sonda ntopng. El profesor, accediendo vía web a dicha sonda, proyecta el mapa de conexiones en tiempo real.

Como prueba inicial, el profesor visita diversas páginas web conocidas ubicadas en Europa (RAE, BBC News, France TV, etc.), verificándose que los servidores efectivamente se encuentran en los países esperados (Fig. 8).



Fig.12. Mapa de ataques Checkpoint



Fig. 13. Advertencia de conexión no segura

Para llevar a cabo un ejercicio práctico más participativo, se solicita la colaboración de todos los asistentes. Se les insta a que, utilizando sus dispositivos móviles personales, realicen las siguientes acciones:

- Activación del "modo avión", de modo que el dispositivo desactive el uso de la red del operador móvil para el envío de datos.
- Conexión a la red Wi-Fi de la sala, utilizando los parámetros de acceso SSID y clave disponibles en cada mesa. Para simplificar este paso, además de textualmente, estos parámetros se facilitan mediante un código QR que permite la conexión directa a la red.
- Uso habitual de sus dispositivos, abriendo las aplicaciones que emplean cotidianamente (correo electrónico, redes sociales, navegación web …).

El mapa de conexiones proyectado en la sala muestra a los asistentes cómo, para su sorpresa, los servidores de dichas aplicaciones se encuentran distribuidos por todo el mundo (Singapur, Rusia, USA, etc.) (Fig. 9).

D. Geolocalización de servicios accedidos en Internet desde dispositivos fijos (conexión al portal cautivo)

Como continuación del ejercicio práctico anterior, se expone al alumnado que, para comprobar las diferencias entre el tráfico que genera un terminal móvil y un terminal de sobremesa, se va a realizar otra prueba, empleando en este caso los equipos de usuario de la sala. Para ello, se solicita que se realicen los siguientes pasos:

- Abran el navegador web usando el icono disponible en el escritorio del SO (previamente arrancado).
- Accedan al servidor de correo electrónico de Gmail usando la URL legítima "https://www.gmail.com/". Para simplificar este paso y maximizar el éxito de la experiencia, la ventana de inicio del navegador ofrece varios iconos de marcadores, entre ellos el de Gmail (Fig. 10).
- Accedan a su cuenta de correo electrónico y consulten sus mensajes, para lo cual cada usuario introduce sus credenciales. En este instante es cuando se produce la captura (anonimizada) del usuario/clave de los asistentes, sin que en general estos advirtiesen ningún comportamiento extraño, a pesar de que las credenciales eran solicitadas dos veces (en primer lugar, por el *servidor de phishing*, y en segundo por el *servidor legítimo*) .



Fig. 14. Advertencia de certificado no firmado por CA de confianza



Fig.15. Conexión sobre certificado firmado por una CA de confianza

Tras ello, el profesor actualiza el mapa de conexiones proyectado, comprobando que en este caso, aunque siguen existiendo conexiones a servicios distribuidos por todo el mundo, el número de conexiones es menor al que se obtenía con los dispositivos móviles.

E. Observación de ciberataques en tiempo real

Como última muestra de la importancia de la ciberseguridad y de los riesgos a los que estamos expuestos en la Red, se muestran dos ejemplos de servicios que ofrecen un mapa mundial sobre el que se representan visualmente ataques que se están produciendo en tiempo real [19, 20] (Fig. 11 y Fig. 12).

F. Muestra de contraseñas capturadas (ofuscadas)

Como etapa final de la actividad, se revelaron las credenciales capturadas, de manera parcial, mostrando solo los primeros tres caracteres del usuario y últimos dos de la clave (Fig. 3). De este modo, se lograba que cada estudiante pudiera reconocer su propio intento de inicio de sesión comprometido, pero respetando su privacidad, al no ser posible para el resto de asistentes extraer las credenciales completas.

G. Explicación simplificada de la técnica de ataque y consejos básicos de seguridad

Para concluir, se explicaron los detalles del proceso de ataque utilizado, junto con una lista de consejos clave para identificar y evitar este tipo de situaciones en un futuro, especialmente cuando se visiten páginas que manejen información personal o delicada (correo electrónico, entidades bancarias, administraciones públicas, etc.).

Como primer consejo, se les indicó que cuando se utilizan dispositivos ajenos que no sean de una entidad de confianza, se debe evitar usar cuentas personales o introducir información confidencial. Si dichos dispositivos han sido manipulados, podemos ser víctimas de multitud de ataques, especialmente del robo de información. Por ejemplo, podemos sufrir fácilmente un ataque de suplantación de páginas web, como ha sucedido en el experimento realizado, sin que tengamos la opción de detectarlo, al haberse alterado los mecanismos de protección de las aplicaciones. Igualmente, se podrían haber instalado en dichos equipos otros programas maliciosos, como registradores de teclas (*keyloggers*) o programas de control remoto, capaces de capturar todas nuestras acciones.

En caso de que el equipo pertenezca a una entidad de confianza, resulta menos probable que haya sido manipulado por el administrador. Sin embargo, si el equipo es de uso compartido entre varios usuarios, el ataque podría provenir de alguno de esos otros. Para minimizar este riesgo, debemos comprobar que el equipo satisface diversos requisitos mínimos de seguridad: tras cada uso se elimina la información de los usuarios y se restablece la configuración original, los usuarios no tienen permisos de administrador (para que no puedan instalar programas maliciosos que afecten al resto de usuarios ni alterar la configuración del sistema) y existe algún programa antivirus o ERP (*Endpoint Protection Platform)*.

Si el dispositivo que estamos utilizando es propio, con garantías de que no haya sido previamente manipulado por un ciberataque, el riesgo es menor, pero no inexistente. En este caso, para que la navegación por Internet sea segura, evitando ser víctima de un ataque de *phishing* como el de la experiencia, deben realizarse varias comprobaciones:

- La conexión es segura, usando el protocolo HTTPS: para ello, basta comprobar que la URL comienza con "https://". Muchos navegadores modernos muestran una advertencia cuando se accede a un sitio web no seguro (Fig. 13).
- El dominio de la URL es completamente correcto: en general, los atacantes no pueden conseguir certificados firmados por una CA de confianza para un dominio que no les pertenece. Así, suelen crear dominios con nombres muy similares al legítimo (por ejemplo, www.ggmail.com", en lugar de "www.gmail.com") para que la diferencia pase desapercibida y engañar al usuario.
- El certificado digital usado por el servidor está firmado por una CA de confianza: si no lo estuviera, el navegador (siempre que no esté manipulado como sucedía en la experiencia) mostraría un mensaje de advertencia de seguridad al cargar la página (Fig. 14), mensaje que el usuario no debería descartar. Cuando el certificado sí está firmado por una CA de confianza, el navegador indica que la conexión es segura mediante un icono de candado cerrado en la barra de navegación, pudiéndose acceder a los datos de la cadena de certificados del servidor haciendo clic sobre dicho icono (Fig. 15).

Por último, se comentó la importancia de activar, siempre que sea posible, la autenticación en dos fases (2FA) para cualquier tipo de servicio que usemos. En caso de que fuese robada nuestra contraseña, este mecanismo evitaría que el ciberdelincuente pueda acceder a nuestra cuenta.

IV. Resultados experimentales

La experiencia fue realizada durante dos días diferentes, a razón de cuatro sesiones cada día. En cada sesión participaron 38 estudiantes de media, con un total de 289 estudiantes.

La gran mayoría del alumnado fue víctima del ataque de suplantación, oscilando el número de credenciales capturadas en cada grupo entre el 85% y el 100% del total de participantes, lo que pone de manifiesto la gran vulnerabilidad a la que están expuestos los usuarios que no están debidamente formados ni concienciados en cuestiones de seguridad. Solamente un pequeño número de ellos fue reticente, con gran acierto, a introducir sus credenciales en un entorno que no les ofrecía las garantías adecuadas. No obstante, dichos usuarios no resultaron víctimas por simple desconfianza, pero ninguno de ellos reparó en comprobar las credenciales del certificado digital del servidor web al que se conectaban (en las que rápidamente hubiesen detectado datos de alerta, como el nombre "Servidor Phishing"), cuestión que pone de manifiesto la necesidad de mejorar la formación en seguridad de los usuarios digitales.

La experiencia tuvo un alto y positivo impacto tanto en los estudiantes como en el profesorado que les acompañaba. Clara prueba de ello fueron las muestras de asombro cuando se proyectaban las credenciales capturadas, los espontáneos aplausos tras cada una de las demostraciones o las felicitaciones recibidas, tanto por parte de los profesores acompañantes, como de la propia dirección de las jornadas tras recibir la realimentación de profesores y estudiantes.

También se puso de manifiesto el alto interés suscitado en los grupos participantes, que centraron más del 80% de sus preguntas en las medidas de prevención necesarias y en los mecanismos de protección a aplicar en sus dispositivos (cámaras, routers, terminales, etc.) para defenderse ante amenazas de ciberseguridad como las mostradas, lo que corroboraba que se había logrado con éxito el objetivo de concienciación buscado con esta experiencia.

V. Conclusiones

En los últimos años, las tecnologías digitales han evolucionado a una velocidad cada vez mayor. Dicho avance supone funcionalidades mejores y más potentes, pero también mayores riesgos de seguridad. Los ciberdelitos se han ido incrementando año tras año. Las mayoría de los usuarios desconocen cómo evitar los peligros de ciberseguridad y, peor aún, que están expuestos a ellos. Uno de los principales pilares para afrontar este problema es invertir en la formación en ciberseguridad de los usuarios.

Este trabajo tiene como objetivo el diseño y realización de una propuesta de formación y educación básica en ciberseguridad. La propuesta ha sido probada con adolescentes, uno de los colectivos con mayor índice de riesgo de sufrir un ciberdelito, dentro de las Jornadas de Puertas Abiertas organizadas por la Escuela Técnica Superior de Ingeniería de la Universidad de Sevilla.

Para llevar a cabo la experiencia, se ha dotado un laboratorio con capacidad de monitorización del tráfico y geolocalización de los servicios, así como de los elementos necesarios para llevar a cabo, con fines didácticos, un ataque de suplantación web o phishing mediante portal cautivo.

Los resultados de la experiencia confirman que muchos de los usuarios de Internet, aunque lo sean desde hace tiempo, como los jóvenes nativos digitales, desconocen el funcionamiento básico de la red y confían con demasiada facilidad en las infraestructuras de acceso y en los dispositivos ajenos. A pesar de los intentos realizados en los últimos años por organismos públicos y privados para concienciar a la población en materia de ciberseguridad, no se observa una mejora sustancial en el conocimiento que poseen la mayoría de los usuarios a la hora de protegerse ante los ciberdelitos. Al menos el 85% del alumnado que participó en el simulacro de phishing no fue consciente del ataque, ni adoptó ninguna medida de precaución básica.

Esta experiencia de formación, con casos prácticos vividos en primera persona, ha supuesto un éxito, tanto por el elevado nivel de participación, como por el interés suscitado en los

jóvenes que han participado en la importancia de la ciberseguridad, lográndose los objetivos buscados. Más del 80% de las preguntas formuladas por los asistentes fueron relativas a las buenas prácticas de seguridad que deben aplicar para evitar ser nuevamente víctimas de ciberdelitos similares. La experiencia, percibida como amena e impactante por el alumnado, fue valorada muy positivamente tanto por los asistentes como por la organización de las jornadas.

El trabajo aporta un posible esquema para desarrollar experiencias de educación que, con una cantidad de trabajo relativamente baja y empleando herramientas de código abierto y gratuitas, ha mostrado excelentes resultados. El trabajo podría reutilizarse en otros centros educativos y universidades para realizar actuaciones de formación en ciberseguridad similares. Asimismo, también podría aplicarse a cualquier grupo de usuarios de Internet (estudiantes o no), al que se desee concienciar o formar en nociones básicas de ciberseguridad. A fin de facilitar la interacción con el profesor y responder dudas de forma ágil, se recomiendan grupos formados por unos 20 participantes.

Como mejora para futuras ediciones, se plantea aumentar ligeramente el tiempo de la experiencia, de modo que los asistentes puedan experimentar como las medidas de mitigación actualmente implementadas en los dispositivos móviles habrían evitado este ataque, tales como HSTS o Certificate Transparency. También se plantea implementar un mecanismo automatizado, basado en encuestas simples o mediante herramientas conocidas como Kahoot, para recoger la opinión de los asistentes una vez finalizada la experiencia, que permita aplicar nuevas mejoras.

AGRADECIMIENTOS

Este trabajo ha sido desarrollado en colaboración con la Cátedra de Ciberseguridad y Ciberinteligecia BeOneSec [21], dentro de las Jornadas de Puertas Abiertas organizadas por la dirección de la Escuela Técnica Superior de Ingeniería de la Universidad de Sevilla. El mantenimiento del laboratorio y el despliegue de la infraestructura necesaria para la experiencia han sido asistidos por el técnico del Departamento de Ingeniería Telemática, Juan José Muñoz González.

REFERENCIAS

[1] Ministerio del Interior (España), "Portal estadístico de criminalidad", https://estadisticasdecriminalidad.ses.mir.es/publico/portalestadistico/ (último acceso 9 febrero 2025).

[2] Ministerio del Interior (España), "Balance de Criminalidad de 2023", https://www.interior.gob.es/opencms/export/sites/default/.galleries/galeria-de-prensa/documentos-y-multimedia/balances-e-informes/2023/Balance-de-Criminalidad-Cuarto-Trimestre-2023.pdf (último acceso 9 febrero 2025).

[3] A. D. Z. Rendón, "Impacto de la inteligencia artificial en los ciberataques", Revista científica Sinapsis (ITSUP), vol. 24, no 1, 2024.

[4] T. Usman, I. Ahmed, A. K. Bashir, K. Shaukat , "A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review", Sensors, vol. 23, no 8, pp. 4117, 2023. https://doi.org/10.3390/s23084117

[5] I. Agrafiotis, J. R. C. Nurse, M. Goldsmith, S. Creese, Upton D, "A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate", Journal of Cybersecurity, vol. 4, no 1, pp. tty006, 2018. https://doi.org/10.1093/cybsec/tyy006

[6] INCIBE. Programa de formación en ciberseguridad, https://www.incibe.es/incibe/formacion (último acceso 27 febrero 2025).

[7] Universidad de Sevilla. Jornadas Preuniversitarias: Formación e Investigación en la Escuela Técnica Superior de Ingeniería de Sevilla, https://etsi.us.es/futuros-estudiantes/jornadas-preuniversitarias (último acceso 10 enero 2025).

[8] T. O. Abrahams, O. A. Farayola, S. Kaggwa, P. U. Uwaoma, A. O. Hassan, S. O. Dawodu, "Cybersecurity awareness and education programs: a review of employee engagement and accountability", Computer Science & IT Research Journal, vol. 5, no 1, pp. 100-119, 2024. https://doi.org/10.51594/csitrj.v5i1.708

[9] S. Hu, C. Hsu, Z. Zhou, "Security Education, Training, and Awareness Programs: Literature Review", Journal of Computer Information Systems, vol. 62, no 4, pp. 752–764, 2021. https://doi.org/10.1080/08874417.2021.1913671

[10] M. Fernández, J. Lizárraga, I. Vélez de Mendizábal, A. Rodríguez, U. Zurutuza, "Guardianes de la Galaxia: concienciación en Ciberseguridad", Actas IX Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2024), Universidad de Sevilla, Escuela Técnica Superior de Ingeniería Informática, pp. 197.203, 2024.

[11] R. Jouaibi, A. K. Gaylard, B. A. Lee, "Employee Cyber-Security Awareness Training (CSAT) programs in Ireland's financial institutions", Cyber Research Conference - Ireland (Cyber-RCI), pp. 1-4, 2022. https://doi.org/10.1109/cyber-rci55324.2022.10032683

[12] H. Alhawas, N. Althemali, R. Alkhaldi, R. Alziyadi, M. Alabdulhafith, "Cybersecurity Training and Awareness System Through Virtual Reality", CS & IT Conference Proceedings, vol. 14, no 5, 2024. https://doi.org/10.5121/csit.2024.140505

[13] The Ntop Team, "Ntopng: High-Speed Web-based Traffic Analysis and Flow Collection", https://github.com/ntop/ntopng/ (último acceso 10 enero 2025).

[14] Eris Networks S.A.S, "DB-IP Lite: Free IP geolocation database", https://db-ip.com/db/lite.php (último acceso 8 febrero 2025).

[15] Google LLC, "Gmail: Servidor de correo electrónico", https://www.gmail.com/ (último acceso 8 enero 2025).

[16] T. Rayat, "Zphisher: A beginners friendly and automated phishing tool", https://github.com/htr-tech/zphisher (último acceso 8 enero 2025).

[17] The Apache Software Foundation, "Apache: HTTP Server Project", https://httpd.apache.org/ (último acceso 8 enero 2025).

[18] Shodan, "Search Engine for the Internet of Everything", https://www.shodan.io/ (último acceso 8 enero 2025).

[19] Kaspersky Labs, "Cyberthreat live Map", https://cybermap.kaspersky.com/ (último acceso 9 febrero 2025).

[20] Check Point Software Technologies, "Live cyber threat map", https://threatmap.checkpoint.com/ (último acceso 9 febrero 2025).

[21] BeOneSec y Universidad de Sevilla, "Cátedra de Ciberseguridad y Ciberinteligecia BeOneSec", https://catedra.us.es/ciberbeonesec/ (último acceso 9 febrero 2025).

Cibersapiens: Gamifying Cybersecurity Education to Engage Secondary Students

Soraya Frías-Guadalupe[1], Pedro Paredes[2], Carlos Castaneda-Marroquín[3]
[1]Facultad de Ciencias de la Educación y del Deporte y Estudios Interdisciplinares
s.friasg@alumnos.urjc.es
[2]Escuela Técnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos
pedro.paredes@urjc.es
[3]Serval Networks, Av. de la Industria, 4, Alcobendas, 28108, Madrid, Spain
ccastaneda@servalnetworks.com

***Abstract*—Cybersecurity education and awareness are often tailored for businesses, leaving general citizens, especially young people, unprepared. Despite the availability of educational resources, cybersecurity training is not provided equally to all individuals, particularly children and teenagers. As digital risks such as sexting, grooming, and cyberbullying, among others, increase, young users remain vulnerable. However, existing awareness efforts fail to engage them effectively and current content lacks the motivation needed to drive meaningful behavioral change. This paper introduces "Cibersapiens: Digital Well-being for a Connected Generation," an interactive, gamified environment designed to teach cybersecurity to secondary school students. The project focuses on developing and implementing gamified educational resources that actively engage students, improve their understanding of cyber risks, and promote best practices in digital security. By leveraging active learning methodologies, this approach transforms cybersecurity education into an engaging and meaningful experience.**

***Index Terms*—Cybersecurity education, Gamification, Student engagement, Active learning, Behavioral change, Cyber awareness**

Contribution type: *Training and educational innovation*

I. Introduction

The increasing digitalisation of daily life has exposed young people to various cybersecurity threats. Despite existing educational resources, cybersecurity awareness among adolescents remains low, as traditional approaches do not engage them effectively [1]. Many young users lack the necessary digital literacy to recognize and respond to online risks, making them particularly vulnerable to manipulation and exploitation [2]. In addition, cybersecurity education is often fragmented and not adapted to the specific needs of teens, leading to a gap between knowledge and practical application. Addressing this issue requires innovative student-centred methods that not only provide theoretical knowledge but also foster behavioral change through engaging and interactive learning experiences [3]. Teenagers aged 12 to 16 are highly active online, yet they often underestimate digital risks [4]. Many do not recognize the consequences of unsafe behavior, making them vulnerable to cyber threats such as identity theft, misinformation, and online harassment [5]. Additionally, their digital interactions increasingly involve social networks, gaming platforms, and online communities, where personal information is frequently shared without appropriate safeguards [6]. Although schools introduce basic digital literacy, these efforts are often insufficient, lacking engaging methodologies that resonate with students. As cyber threats evolve, there is a growing need for cybersecurity education that is interactive, relevant, and capable of fostering long-term behavioral change [7].

Gamification is an educational approach that incorporates game mechanics to enhance learning experiences [8]. Research has shown that it increases motivation, engagement, and retention, making it a valuable tool in education [9]. Applied to cybersecurity training, gamification can create immersive learning environments that help students develop awareness and digital security skills while making the learning process more enjoyable and interactive. Traditional cybersecurity education often relies on passive methods, such as lectures and informational materials, which do not capture the students' attention or encourage long-term behavioral change [10].

For secondary school students, particularly those aged 12 to 16, gamification is especially effective as it aligns with their natural inclination toward interactive and competitive activities. By embedding cybersecurity concepts into engaging challenges, role-playing exercises, and simulated cyber incidents, students develop a deeper understanding of online risks, such as phishing, social engineering, and privacy breaches. Ultimately, gamification transforms cybersecurity education into a meaningful and impactful experience that prepares students to navigate the digital world responsibly and securely.

This paper presents Cibersapiens: Digital Well-being for a Connected Generation, a gamified educational framework designed to teach cybersecurity to secondary school students. The project aims to address the significant gap in cybersecurity education for young people by integrating interactive, game-based learning into their daily educational experience.

Through Cibersapiens, our aim is to transform the way cybersecurity is taught by moving away from passive, lecture-based instruction to active, hands-on learning. This transition is crucial because young people are increasingly exposed to digital dangers and traditional educational tools often do not prepare them adequately. By incorporating gamification, Cibersapiens not only enhances students' technical knowledge but also instils lasting behavioral changes that promote respon-

sible and secure digital practices. Ultimately, this paper highlights the transformative potential of gamification in fostering cybersecurity awareness and skills among the next generation of Internet users.

The main contributions of this paper are summarized as follows:

- A discussion on best practices for integrating gamification into cybersecurity education: This paper identifies and discusses effective strategies for integrating gamification into the curriculum. Drawing on lessons from Cibersapiens, we provide guidelines and recommendations on how to design gamified cybersecurity programmes that resonate with students, increase their engagement, and address specific challenges they face in the digital world. By outlining these best practices, the paper aims to support educators in adopting gamification as a powerful tool for teaching cybersecurity.
- An alignment of gamified cybersecurity education with national and European competence frameworks: This paper presents a structured review of how gamified cybersecurity education can align with the digital competence goals defined by the Spanish LOMLOE education law and the European DigComp framework. This ensures that the proposed activities are not only engaging, but also pedagogically sound and consistent with educational standards. By connecting learning outcomes with specific competence areas, the paper supports the integration of cybersecurity education into broader objectives like digital literacy, responsible use of technology, and the development of critical thinking.
- Design and implementation of Cibersapiens, a gamified cybersecurity learning environment: This paper presents the development and execution of Cibersapiens, an innovative and interactive platform that uses gamification to teach cybersecurity to secondary school students. The design focusses on creating an engaging and immersive learning environment where students can actively participate in realistic cybersecurity challenges and scenarios, helping them better understand digital risks and best practices for online security. In adittion, the implementation of Cibersapiens takes into account the participation of the entire educational community, including teachers, parents, and school administrators, ensuring a holistic approach to fostering digital well-being and cybersecurity awareness.

This paper is structured as follows. Section II provides an overview of effective ways to increase awareness of digital risks and emphasizes the importance of addressing these issues early in students' lives. III reviews existing literature on cybersecurity awareness in secondary education, highlighting the challenges and current approaches to teaching digital security. IV, introduces the design and implementation of Cibersapiens, detailing the educational principles and gamification techniques applied on the platform. Section V, presents initial feedback from educators and students, offering valuable insight into the effectiveness and engagement of the gamified approach. Finally, Section VI, Conclusions and Future Work, summarizes the findings of the study and outlines directions for future research, focussing on refining the platform and expanding its application to other educational contexts.

II. Motivation

The teaching of cybersecurity to adolescents aged 12 to 16 is motivated by several crucial factors. At this age, young people are highly connected to the digital world. Understanding cybersecurity empowers them to protect their personal information, avoid identity theft, and mitigate risks such as cyberbullying or online fraud [11]. In addition, training young people in this area promotes responsible digital citizenship by fostering an understanding of the impact of their online actions and promoting respect for the privacy and rights of others [12].

Educating adolescents about cybersecurity also indirectly contributes to a more secure digital ecosystem. Reducing vulnerable behaviours minimises the likelihood of cyberattacks on a community or even global scale [7]. Moreover, digital literacy is essential for future careers, not only in technology-related fields, but in all sectors of an increasingly digital society [13]. Cybersecurity education encourages risk analysis, critical thinking, evaluation of information sources, and informed decision-making skills that extend far beyond the digital realm.

Despite these benefits, the exponential growth in the use of digital technologies by adolescents has not been accompanied by a parallel increase in cybersecurity awareness or digital responsibility. Although cybersecurity-related content has been formally included in the curriculum of the Community of Madrid, studies reveal that adolescents remain highly vulnerable to online risks such as cyberbullying, identity theft, sextortion, and misinformation. Traditional educational approaches, often based on theoretical sessions or isolated workshops, do not engage students effectively or promote long-term behavioural change.

Recent literature highlights the scarcity of comprehensive evidence-based cybersecurity education programs. Among the few that exist, even fewer implement active learning methodologies or gamification, despite the documented benefits of these strategies in terms of student motivation, meaningful learning, and knowledge retention. Additionally, the lack of programs targeting both students and their families represents a critical gap, considering that adolescents often depend on adults who themselves lack sufficient training in safe digital practices.

This situation calls for the design and implementation of innovative, dynamic and empirically grounded educational interventions capable of reaching young learners in a language and format that resonate with their interests and lived experiences. Gamification, as an active pedagogical approach, offers a powerful means of turning cybersecurity education into an engaging and impactful experience. In this context, Cibersapiens emerges as a novel educational proposal that integrates digital citizenship, cybersecurity, and game-based

learning to empower secondary school students to make safe, ethical, and responsible decisions in the digital world.

III. State of the art

A. Gamification

Gamification involves the application of game mechanics, elements, and design techniques in non-game contexts to engage users and address problems.[14] [15]. This active methodology has experienced significant growth in recent years, placing the student at the center of the learning process in an autonomous and participatory manner. [16]. Active teaching methodologies are grounded in the constructivist pedagogical model, providing an appealing alternative to traditional education by placing greater emphasis on what the student learns rather than what the teacher instructs. This approach fosters deeper understanding, increased motivation, and greater participation of the student in the learning process. [17]

Several authors have identified the main advantages of this methodology, highlighting an increase in student motivation, the creation of a safe learning environment, increased student awareness of their progress, the promotion of cooperation, the improvement of self-awareness with respect to abilities and the enhancement of retention of acquired knowledge in a meaningful way.[9]

Gamification emerges as one of the most effective methodologies for teaching cybersecurity content, outperforming traditional methods currently in use, which have shown limited impact among adolescents. In this regard, the study "Leveraging Gamification and Game-based Learning in Cybersecurity Education: Engaging and Inspiring Non-Cyber Students" highlights the suitability of gamification to engage and inspire students who are not specialized in cybersecurity. Practical experiences, such as "Capture the Flag" (CTF) events, are particularly effective as they enable students to solve problems in a dynamic and appealing environment. [10]

B. Cybersecurity Education in the Curriculum of the Community of Madrid

In the Community of Madrid, the subject "Ciencias de la Computación", an elective in 1st and 2nd year of ESO, develops specific competency 4: understanding how computer systems communicate to form networks and the importance of cybersecurity.

- 1st ESO: Safe and responsible use of the Internet.
- 2nd ESO: Cybersecurity (Internet security, types of attacks, threat identification, and protection software).

The subject "Tecnología y Digitalización", mandatory in 2nd and 3rd year of ESO, focuses on:

- Competency 6: Understanding how devices and applications work, optimizing their use, and solving technical issues.
- Competency 7: Using technology responsibly, assessing its impact.

In 2nd ESO, the digitalization of the personal learning environment is covered:

- Safe Internet use: information search, email, messaging, and social networks.
- Learning tools and platforms: configuration, maintenance, and critical use.
- Network security: risks, threats, attacks, and protection measures (antivirus, firewalls, proxy).
- Good digital practices: cyberbullying, sextortion, privacy, inappropriate content, and addiction risks.

In 4th ESO, the subject "Digitalización" (Technological option) includes:

- Competency 3: Digital well-being habits, device and data protection, and health.
- Competency 4: Critical digital citizenship and responsible use of technology.

The following topics are covered:

- Digital security and well-being: Introduction to cybersecurity, device security (antivirus, backups), data protection (digital identity, privacy, GDPR), and health risks (ergonomics, mental well-being).
- Critical digital citizenship: Freedom of expression and its limits, netiquette, intellectual property, digital journalism, fake news detection, online administrative procedures (electronic ID, digital signature), secure e-commerce, and ethics in data usage (AI, planned obsolescence, digital sustainability).

C. Cybersecurity Education Programs

Training in digital skills has not kept pace with the rapid technological development, leading to significant gaps in adolescents' knowledge about how to protect themselves in digital environments. Despite the increasing use of technology, education about its safe and responsible use has not progressed at the same rate. As a result, the responsibilities associated with the intensive use of ICT are not adequately addressed [18].

Despite the wide availability of cybersecurity content in high-quality repositories, such as those offered by INCIBE, the real gap lies in the lack of effective educational proposals that use methodologies capable of making a real impact on students and helping them identify digital risks. One-off initiatives, such as workshops or informational talks, often lack the appeal and effectiveness of more dynamic and interactive approaches, such as gamification and game-based learning.

To assess the current state of cybersecurity education, searches were performed in academic databases such as Dialnet, Scopus, and ERIC, as well as in search engines such as Google Scholar. The goal was to identify studies that focus on preventive actions implemented in educational settings. However, the number of relevant results was limited and increased further when filters related to innovative methodologies were applied.

This shortage of programs promoting responsible Internet use and digital risk management is reflected in the study "Responsible Internet Use and Digital Security: A Systematic Review of Educational Programs". This research highlights the

limited number of published empirical studies that examine existing educational programs in this area and evaluate their results. Of a total of 702 articles initially considered, only 19 were selected for relevance. Although international organizations have proposed initiatives to address digital risks, very few educational programs have been developed on the topic. Furthermore, the scientific literature stresses the urgent need to design and implement training programs on digital security and responsible technology use aimed at the entire educational community [19].

The article highlights that cybersecurity education programs have been implemented primarily in Spain (36.8%) and the United States (15.8%). Of these, 68.4% targeted students, mainly children and young people up to 13 years of age. Most of the programs were sort, usually lasting a few weeks, with only a few designed as continuous interventions. Common topics addressed included cyberbullying, sexting, grooming, identity theft, and misinformation.

Among the 19 programs analyzed, only one incorporated gamification as a strategy to evaluate learning about privacy concepts. It is particularly concerning that none of these programs were aimed at families, despite the fact that children often rely on adults for guidance on safe Internet use, although many of these adults lack the necessary training themselves [19].

In this regard, Prats Fernández et al.[3] emphasize that, although there have been efforts to develop programs aimed at promoting responsible use of ICT, their implementation in real educational contexts remains limited. This has resulted in a scarcity of publications evaluating the effectiveness of preventive measures and interventions related to ICT-related risks. The authors stress the importance of continuing to promote studies that provide relevant data on the development and impact of such measures for adolescents. Furthermore, they highlight a clear need and social demand for ongoing education in this area to raise awareness and reduce the risks associated with technology.

It is essential to develop meaningful work in cyber education to help preadolescents understand the risks associated with the digital environment. The creation of interactive and gamified environments emerges as one of the most effective strategies to teach cybersecurity, with a purpose that goes beyond simple awareness. The goal is to design training programs that encourage children to adopt good Internet security practices, transforming them into habits through a process of assimilation that integrates technological actions into their daily behavior through continuous practice.

IV. Method

A. Educational intervention

Cibersapiens: Digital Wellbeing for a Connected Generation, is an interactive and gamified environment aimed at enhancing cybersecurity education at the lower secondary level. This initiative focuses on the design and implementation of gamified instructional resources that foster student engagement and a deeper understanding of cyber risks and responsible digital practices, through the application of active learning methodologies. The overarching goal is to transform cybersecurity instruction into an engaging and meaningful educational experience that motivates students to take an active role in their learning process, while fostering greater awareness and long-term involvement in the entire educational community.

B. Contents

The content of the proposal is structured around three fundamental pillars: conceptual, procedural, and attitudinal. The first module, "Introduction to Cybersecurity," introduces key cybersecurity concepts and highlights the importance of protecting personal information in digital environments. The second module, "Risks in the Digital Environment," focusses on the procedures needed to identify and manage the main risks faced by adolescents, including cyberbullying, grooming, sexting, loss of privacy, and security threats such as malware, phishing, and ransomware. This module also addresses the impact of disinformation, how to detect fake news, and the symptoms associated with technology addiction, particularly its consequences for mental and physical health, along with strategies for achieving a healthy and balanced use of technology. Finally, the third module, "Prevention and Protection Strategies," emphasizes the importance of fostering responsible and ethical attitudes in the use of technology, encouraging critical and reflective thinking in the face of digital challenges. Covers the responsible use of social networks, privacy settings, how to handle risky situations (including support resources and who to turn to), and offers practical tools to begin learning about cybersecurity.

Session	Thematic Block	Content
1	Introduction to cybersecurity	Key concepts and their importance
2	Risks in the digital environment	Cyberbullying, grooming, and online harassment
3	Risks in the digital environment	Sexting and loss of privacy
4	Risks in the digital environment	Security threats (malware, phishing, ransomware) and disinformation
5	Risks in the digital environment	Technology addiction and its impact on physical and mental health
6	Prevention and protection strategies	Cybersecurity education and privacy settings
7	Prevention and protection strategies	Responsible use of social networks and managing risky situations
8	Wrap-up and assessment	Final reflection and learning evaluation

Table I: Session plan for the cybersecurity education program

As you can see in table I, the content of the program is organized into eight 50-minute sessions, aligned with the standard school schedule. These sessions are structured around three thematic blocks: (1) Introduction to cybersecurity, (2) Risks

in the digital environment, and (3) Prevention and protection strategies. This organization enables a progressive exploration of the conceptual, procedural, and attitudinal dimensions of cybersecurity education. The detailed session plan is described below:

Through the proposed content, the program promotes the development of key competencies aligned with the Secondary Education curriculum (ESO), following Decree 65/2022 of July 20, issued by the Governing Council, which establishes the organization and curriculum for this educational stage in the Community of Madrid. In particular, this intervention enables students to strengthen essential competencies such as digital competence, learning to learn, citizenship (social and civic) competence, and linguistic communication skills. These capabilities are crucial for students to communicate effectively in digital environments and to become active and responsible citizens in today's society.

This content proposal is also closely aligned with the Digital Competence Framework for Citizens (DIGCOMP), an essential European Union-level tool for fostering digital competence among citizens. Specifically, it focusses on the areas of security (area 4) and problem solving (area 5), aiming to equip students with the ability to protect their devices, content, personal data, and privacy in online environments. Furthermore, it seeks to raise awareness of physical and mental health about the use of technology and provide the skills necessary to identify needs and solve problems in digital contexts.

The primary goal is for students to develop these competencies by progressing from a basic level with teacher guidance to a basic level with autonomy, while still receiving support when needed. The initiative aims to provide students with educational materials that promote an initial level of cybersecurity awareness. Although valuable repositories such as the one offered by the Spanish National Cybersecurity Institute (INCIBE) can be used as complementary resources, this intervention is distinguished by its emphasis on a formative rather than merely informative approach. Through the combination of educational materials and active methodologies such as gamification, the initiative seeks to generate a meaningful impact on students, promoting an initial level of awareness of the responsible and safe use of technology.

In this regard, a recurring educational design is proposed for schools, one that involves not only students but also the broader educational community, with particular emphasis on families and teaching staff. This comprehensive approach aims to foster a culture of cybersecurity awareness and best practices that go beyond the classroom and reach the social environment of students.

With regard to teachers, the contents of this intervention proposal are designed to support their training in digital competencies. These materials are expected to help teachers reach the A2 proficiency level, as well as raise their awareness of the importance of these skills, enabling them to effectively transfer this knowledge to their students during their lessons.

Likewise, the training proposal extends its reach to students' families, recognizing their essential role in the development of digital competencies. The content will provide families with clear information and practical tools to understand the risks and opportunities inherent in the digital world. In this way, the objective is to empower parents and guardians to actively support their children to acquire safe, responsible, and critical habits in the use of technology, fostering a family environment that promotes positive digital citizenship.

In the CiberSapiens: Digital Wellbeing for a Connected Generation proposal, gamification stands out as the main methodological strategy for teaching cybersecurity due to its ability to increase student motivation and engagement. By incorporating game elements into the teaching-learning process, students are expected to feel more interested and willing to participate in cybersecurity-related activities. Furthermore, gamification allows complex topics to be addressed in a playful way, facilitating both understanding and information retention.

This proposal also promotes active learning and student engagement through practical activities, debates, and collaborative projects. Encourage teamwork, critical thinking, and cooperation among students to solve problems and share knowledge, fostering cooperative learning.

C. Materials and Platforms

This intervention proposal relies on a variety of resources aimed at maximizing adolescent engagement and understanding in the field of digital security.

First, it is essential to have adequate technological equipment, such as desktop computers, laptops, and tablets, that allows students to access content and actively participate in the proposed activities. In addition, it is crucial to ensure high-speed internet connectivity in classrooms and/or shared spaces where sessions take place, thus providing a seamless and uninterrupted experience. It is also important to have a projector and screen for theoretical presentations, as well as classrooms equipped with appropriate furniture to support both group and individual activities.

To complement face-to-face activities, the use of a virtual learning platform is proposed, such as EducaMadrid. This platform would function as a repository for materials, a discussion forum, and a tool to monitor student progress. It allows students to access content anytime and anywhere, thus promoting autonomous learning and peer collaboration.

With regard to teaching materials, the creation of attractive and interactive resources is supported by tools such as Canva and Genially. These resources include presentations, infographics, videos, and gamified activities that allow students to learn in an engaging and participatory way. It is also recommended that you have stationery and office supplies available for the creation of informational posters and other supporting materials.

Gamification plays a central role in this proposal, with tools like Educaplay suggested to design engaging activities and interactive assessments that motivate students to learn and apply cybersecurity concepts. These tools allow the creation of games, quizzes and challenges customized to the level of

competence of the students, thus encouraging personalized learning.

Moreover, communication and collaboration tools such as Wikis and Padlet are recommended to facilitate interaction among members of the educational community and collaborative work on cybersecurity-related projects. These tools encourage students to share ideas, resources, and experiences, contributing to a collective understanding of digital security.

To support the development of sessions, specific games have been designed to encourage active participation and practical learning (see table II). These games, which are described below, are integral to the proposal and essential to achieving awareness and learning goals. Cybersecurity Word Roulette, Crossword Puzzle, Role-Playing Game or Escape Room, Digital Detectives, Ciberquiz: Defend Your Digital Space, Privacy Puzzle: Sequencing Strategies, Interactive Map: Exploring Threats and Misinformation, Froggy Challenge on Tech Addiction, Practical Workshop: Configure Your Social Media for Safe Browsing, and Matching Game: Safety in the Digital World.

Lastly, access to Open Educational Resources (OER), such as those available on Procomún and other repositories, as well as the resources provided by INCIBE, is recommended. These materials offer high-quality curriculum-aligned content suitable for various educational levels and contexts, allowing teachers to enrich their sessions and students to deepen their knowledge in areas of personal interest.

Session	**Game**
1. Introduction to cybersecurity	a. Word Roulette (pre-test) b. Crossword (pre-test)
2. Risks in the digital environment	a. Escape Room b. Cyberquiz
3. Sexting and loss of privacy	a. Sequence ordering game
4. Security threats and Disinformation	a. Interactive map
5. Technology addiction	a. Froggy Challenge
6. Cybersecurity education and privacy settings	a. Practical workshop
7. Responsible use of social media	a. Matching situations
8. Closing and evaluation	a. Word Roulette (post-test) b. Crossword (post-test)

Table II: Session and Game Table

D. Sample Activity

Below is a sample gamified activity designed for Session 2, “Cibersapiens in Action: Defend Your Digital Space from Grooming and Cyberbullying!” This session aims to address the risks associated with cyberbullying and grooming, raise student awareness, and promote active learning and critical reflection.

The session begins with a brief introduction that outlines its objectives, followed by a theoretical presentation. In this section, the concepts of cyberbullying, grooming, and online harassment are introduced. The various ways in which these phenomena manifest, their consequences, and prevention strategies are discussed.

Next, the practical part of the session begins, focussing on the game “Digital Detectives.” This activity allows students to explore the topics in a playful and participatory manner, promoting empathy and responsibility in the digital environment.

In “Digital Detectives,” students take on the role of rookie detectives with the mission of rising through the ranks to become professional digital agents. Using the Genially platform, students can access an interactive map that presents four critical missions. In each mission, they face questions and scenarios that require identifying risks, making informed decisions, and applying prevention strategies. Upon completing each mission, students receive clues that allow them to progress and, ultimately, unlock a virtual briefcase by cracking a secret combination.

The first students to solve the puzzle are rewarded with the ”Cybersecurity Kit,” which includes educational resources, practical guides, and a “Professional Digital Agent” diploma — a symbolic recognition celebrating their achievement in the activity. This dynamic not only encourages learning about the dangers of cyberspace but also promotes critical thinking and teamwork.

The main goal of this game is to help students identify and better understand the dynamics of cyberbullying and grooming, as well as to develop skills to make informed and responsible decisions when facing these situations. Through this role-playing game, students gain practical insight into how these threats operate in the digital environment, what warning signs to look out for, and what actions they can take to protect themselves and others.

The activity fosters empathy by placing students in the position of victims and encouraging reflection on the consequences of their actions in the digital world. By assuming the role of detectives, students must analyze complex scenarios, assess risks, and make decisions that affect the well-being of others, helping them develop greater awareness of the importance of acting responsibly and compassionately in cyberspace.

After the “Digital Detectives” activity, the students participate in a second game: “Cyberquiz: Defend Your Digital Space.” In this activity, students access an interactive online quiz featuring definitions and questions related to grooming, cyberbullying, and online harassment. Working in teams, they participate in a competitive quiz designed to assess their understanding of the concepts covered. In the end, the answers are reviewed, and key definitions and ideas are discussed, including topics such as grooming, cyberbullying, sexting, and data privacy, which encourage teamwork and collaboration. The goal of this activity is to reinforce the students’ understanding of the key terms and concepts related to online abuse, grooming, and cyberbullying.

Finally, the session concludes with a group reflection in which the students share their thoughts on cyberbullying and discuss possible intervention strategies, thus consolidating what they have learnt during the session. In conclusion, both “Digital Detectives” and “Cyberquiz” serve as playful and

effective tools to raise awareness among students about the dangers of cyberbullying and grooming, provide them with practical strategies to prevent and address these issues, and promote a culture of respect and responsibility in the digital world.

E. Engaging the Educational Community

Cybersecurity is a crucial aspect of adolescent education, and its resolution requires active collaboration of the entire educational community, including families and teachers. This intervention proposal aims to create a learning environment in which parents, educators, and experts work together to protect young people in the digital world. Families and teachers are key figures in adolescents' daily lives, and their influence can be decisive in shaping healthy digital habits. Therefore, it both groups must have access to tools and resources related to cybersecurity. Teachers can provide effective training, while families can encourage responsible use of technology at home.

To support this collaboration, a blog titled "Cibersapiens: The Cybersecurity Blog for Parents and Educators" has been created. This platform aims to provide clear, concise, and practical information on online risks and strategies to prevent them. The blog empowers families and teachers with the knowledge they need to guide teens in the safe and responsible use of technology. It is intended for parents, teachers, and educators who care about the well-being of adolescents in the digital environment. The blog includes content designed for students, as well as additional materials, such as infographics, videos, and relevant news.

Among the resources offered are articles on the most common risks in the digital environment, practical guides to configure privacy on social networks and apps, tips for encouraging a balanced use of technology, educational resources and tools to dive deeper into cybersecurity, best practices and examples of how other parents and teachers are addressing cybersecurity challenges, a discussion forum to share experiences and advice, information about events, workshops, and courses, and contact information for cybersecurity experts and support resources.

In addition to the blog, several activities are proposed within the school to engage the entire educational community. These include awareness campaigns, using posters, leaflets, and videos to raise awareness of online risks and encourage responsible technology use; events such as talks, workshops, and exhibitions aimed at students, families, and teachers, delivered by cybersecurity experts; and the creation of a cybersecurity policy, establishing clear rules and procedures to protect the data and privacy of all members of the educational community.

Active participation of families and teachers in cybersecurity education is essential to build a safe and responsible digital environment. Through collaboration and access to appropriate resources, adolescents can be empowered to navigate the digital world in a safe, ethical, and mindful manner.

F. Evaluation Design

The evaluation of the intervention will be structured into three phases: initial, procedural, and final, using a combination of quantitative and qualitative methods. Conceptual learning will be assessed through statistical analysis of data from the Genially and Educaplay platforms, based on students' performance in gamified activities specifically designed. During the initial phase, a pre-test will be administered using a roulette-style game and a crossword puzzle focused on core cybersecurity concepts. These same activities will be repeated in the final phase as a post-test, enabling comparison and measurement of knowledge gains.

Practical and attitudinal learning will be evaluated through direct observation, supported by a structured checklist and a learning diary maintained by the students, in which they will document their experiences, reflections, and observed behaviors throughout the intervention. Finally, a satisfaction questionnaire will be administered to students, along with a separate survey to teachers, aimed at collecting feedback on the perceived effectiveness of the intervention and identifying areas for improvement.

V. Insights and Preliminary Findings

The preliminary implementation of Cibersapiens, a gamified cybersecurity learning environment designed for secondary education, has provided valuable insight into both the current state of digital security education and the potential for innovative pedagogical interventions. Although full deployment and data collection phases are ongoing, early findings provide a strong foundation for reflection, refinement, and future scalability.

A key insight concerns the limited and fragmented presence of cybersecurity topics within the official curriculum. Digital security is often treated as a peripheral subject, typically addressed within optional or highly specialized courses such as Technology or Computer Science, and not systematically integrated into core subjects. In many schools, these topics appear only in the second or third year of compulsory secondary education (ESO), if at all. This curricular gap creates a scenario in which many students complete their formal education with minimal exposure to essential knowledge and practices related to cybersecurity.

Moreover, a review of existing literature and initial field observations highlights a recurring shortcoming in previous initiatives: the lack of continuity, contextualization, and pedagogical depth. Many programs rely heavily on isolated workshops or awareness days, which, while useful, are insufficient to instill sustained behavioral change or deep understanding. Cibersapiens addresses this challenge by proposing a more holistic and embedded model, tightly aligned with curricular standards, and tailored to the developmental needs of adolescents.

The proposed approach combines three interrelated components: (1) a gamified learning environment; (2) a collaborative blog for the educational community; and (3) a school-wide

awareness campaign. The use of gamification—through role-playing scenarios, quizzes, and interactive challenges—has emerged as a particularly promising strategy. These tools engage students in active, contextual learning, encouraging ethical reflection and internalization of safe digital habits. Narrative design and interactivity have proven to be effective in increasing motivation and supporting knowledge retention.

The initial pilot of these gamified activities suggests that interactivity and narrative design significantly enhance student motivation. Role-playing exercises, for instance, have allowed students to explore multiple perspectives (e.g. hacker, victim, bystander), promoting empathy and a deeper understanding of the implications of digital behavior. Similarly, quizzes and minigames have been effective in consolidating technical concepts such as password hygiene, phishing detection, and privacy settings.These learning activities have been well-received by both students and teachers, according to initial qualitative feedback collected through observations and post-session discussions.

One of the most important insights so far is the need to broaden the focus of cybersecurity education beyond the student body. Families and educators play a fundamental role in reinforcing safe practices and supporting students' learning processes. The inclusion of parents and teachers in the initiative through a dedicated blog titled Cibersapiens: A Cybersecurity Blog for Parents and Educators has been designed to build a shared knowledge base and foster a culture of digital responsibility across the entire school community.

The blog acts as both a resource hub and a participatory platform. It provides access to articles, infographics, tutorials, videos, and links to external resources. It also encourages educators and parents to contribute, turning them from passive recipients of information into active agents of cybersecurity education. In its future development, the blog is envisioned as a wiki-style collaborative space, allowing co-creation of content and fostering stronger community ownership of digital safety initiatives. These tools are expected to support the continuous refinement of both content and delivery methods. Nevertheless, sustaining the relevance of content in the face of emerging digital threats will require regular updates and institutional commitment.

Although the project is still in its early implementation stages, another key insight has emerged around the need for robust evaluation tools. To assess the impact of the intervention, it is essential to go beyond knowledge tests and incorporate multidimensional strategies capable of capturing behavioral change, self-efficacy, and perceived risk.

Future phases of the project will prioritize the development of comprehensive assessment instruments, including pre- and postintervention questionnaires, classroom observations, and analytics collected through the gamified platform (e.g., time on task, engagement patterns, decision-making paths). These tools are expected to support the continuous refinement of both content and delivery methods.

Furthermore, feedback loops involving students and teachers are being established to ensure that materials are adaptable to different learning contexts and that activities are perceived as meaningful and culturally relevant.

Another crucial insight relates to the pivotal role of teachers in the success of cybersecurity education initiatives. Although Cibersapiens includes materials and activities aimed at the school community, preliminary observations suggest that many educators feel underprepared to address cybersecurity topics with confidence. This is in part due to the lack of formal training in digital safety and the absence of structured professional development opportunities in this area. As such, empowering teachers through targeted training, ongoing support, and collaborative resource design is essential. Their active involvement not only ensures the continuity and contextualization of learning experiences but also helps foster a school culture where cybersecurity becomes a shared responsibility.

In summary, the preliminary findings of the Cibersapiens initiative underscore the importance of an integrated, participatory, and pedagogically grounded approach to cybersecurity education. The combination of gamified learning, community engagement, and continuous feedback mechanisms appears to offer a promising path forward. These early results will guide the refinement of the intervention and lay the groundwork for its broader application in secondary schools.

VI. Conclusions and Future Work

The development of Cibersapiens highlights the potential of gamified environments as a meaningful response to the growing need for cybersecurity education in secondary schools. By aligning with active learning methodologies and involving the entire educational community, the project fosters both student engagement and collective digital awareness.

Preliminary observations suggest that integrating cybersecurity into the curriculum through experiential and interactive formats can effectively promote critical reflection, responsible online behavior, and long-term attitude change. Furthermore, the project underscores the importance of inclusive strategies that involve not only students but also educators and families.

Future work will focus on refining the proposed activities, expanding the blog into a collaborative platform, and implementing robust evaluation tools to assess learning outcomes and behavioral impact. These steps will contribute to shaping a scalable and adaptable model for cybersecurity education in schools.

Acknowledgment

This work is part of the project I+D+i PID2022-137849OB-I00 and has been supported by MICIU/AEI/10.13039/501100011033/ y FEDER/UE

References

[1] N. A. A. Rahman, I. H. Sairi, N. A. M. Zizi, and F. Khalid, "The importance of cybersecurity education in school," *International Journal of Information and Education Technology*, vol. 10, no. 5, pp. 378–382, 2020.

[2] R. G. Gomez, P. A. Llorente, M. T. V. Morales, and I. L. Hernández, "Seguridad y protección digital de la infancia: retos de la escuela del siglo xxi," *Educar*, vol. 56, no. 1, pp. 219–237, 2020.

[3] M. A. Prats Fernández, A. Torres-Rodríguez, U. Oberst, X. Carbonell *et al.*, "Diseño y aplicación de talleres educativos para el uso saludable de internet y redes sociales en la adolescencia: descripción de un estudio piloto," *Píxel-Bit. Revista de Medios y Educación, 2018, Núm. 52*, 2018.
[4] S. Martín Pacheco *et al.*, "Percepción de los riesgos de internet y las redes sociales en la adolescencia. revisión teórica," 2022.
[5] A. A. Al Shamsi, "Effectiveness of cyber security awareness program for young children: A case study in uae," *Int. J. Inf. Technol. Lang. Stud*, vol. 3, no. 2, pp. 8–29, 2019.
[6] A. G. Jiménez *et al.*, "Una perspectiva sobre los riesgos y usos de internet en la adolescencia," *Revista ICONO 14. Revista científica de Comunicación y Tecnologías emergentes*, vol. 9, no. 3, pp. 410–425, 2011.
[7] M. Bada, A. M. Sasse, and J. R. Nurse, "Cyber security awareness campaigns: Why do they fail to change behaviour?" *International Conference on Cyber Security for Sustainable Society*, 2019.
[8] Y.-k. Chou, *Actionable gamification: Beyond points, badges, and leaderboards*. Packt Publishing Ltd, 2019.
[9] J. Roa González, A. Sánchez Sánchez, and N. Sánchez Sánchez, "Evaluación de la implantación de la gamificación como metodología activa en la educación secundaria española," *REIDOCREA*, vol. 10, no. 12, pp. 1–9, 2021.
[10] L. Williams, E. Anthi, Y. Cherdantseva, and A. Javed, "Leveraging gamification and game-based learning in cybersecurity education: Engaging and inspiring non-cyber students," *Journal of The Colloquium for Information Systems Security Education*, vol. 11, no. 1, 2024.
[11] S. Livingstone, K. Ólafsson, E. J. Helsper, F. Lupi´ñez-Villanueva, G. A. Veltri, and F. Folkvord, "Maximizing opportunities and minimizing risks for children online: The role of digital skills in emerging strategies of parental mediation," *Journal of Communication*, vol. 67, no. 1, pp. 82–105, 2017.
[12] M. Ribble and D. Fodeman, *Digital Citizenship in Schools*. International Society for Technology in Education, 2011.
[13] *OECD Skills Strategy 2019: Skills to Shape a Better Future*, OECD, 2019.
[14] G. Zichermann and C. Cunningham, *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*. O'Reilly Media, Inc., 2011.
[15] K. Werbach and D. Hunter, *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton School Press, 2012.
[16] M. A. Jara Ramírez, M. M. Velásquez Aguirre, W. H. Arboleda Mazo, J. F. Niño Manrique, N. M. Bolaños Tegue, O. L. Mejía Arango, D. Toro Contreras, A. Y. Agudelo Serpa, A. C. Jiménez Ortiz, A. C. Zúñiga Zapata *et al.*, "Estrategias pedagógicas innovadoras," 2022.
[17] M. D. Taipe, "Metodologías activas en el proceso enseñanza-aprendizaje.(revisión)," *Roca: Revista Científico-Educaciones de la provincia de Granma*, vol. 16, no. 1, pp. 463–472, 2020.
[18] P. M. Yellowlees and S. Marks, "Problematic internet use or internet addiction?" *Computers in human behavior*, vol. 23, no. 3, pp. 1447–1453, 2007.
[19] B. Aznar-Martínez, A. Casarramona-Basany, J. Grané-Morcillo, J. Lorente-de Sanz, M.-À. Prats-Fernández, and L. Ballester-Brage, "Uso responsable de internet y seguridad digital: revisión sistemática de programas educativos," *Estudios Sobre Educacion*, 2024.

Gamificación con Genially para aprender criptografía post-cuántica

Édgar Pérez-Ramos, Candelaria Hernández-Goya, Pino Caballero-Gil
Universidad de La Laguna
Tenerife, España
{eperezra, mchgoya, pcaballe}@ull.edu.es

***Resumen*—El desarrollo y la amenaza que representa la computación cuántica plantean grandes retos para la ciberseguridad. En este contexto, la criptografía post-cuántica surge como un campo de investigación crucial. Sin embargo, uno de los principales desafíos que enfrenta este campo es la escasez de recursos didácticos disponibles. El presente trabajo aborda esta problemática mediante la creación de recursos didácticos centrados en los retículos, el problema de aprendizaje con errores y los algoritmos estandarizados por el NIST, utilizando Genially. Para ello, se ha diseñado una actividad basada en gamificación que introduce y relaciona los retículos con diversos conceptos matemáticos impartidos en el aula. La actividad desarrollada es totalmente novedosa y tiene como objetivo hacer más accesible la comprensión de la criptografía post-cuántica mediante herramientas visuales y manipulables, como archivos en GeoGebra y calculadoras dinámicas, que ayuden a los estudiantes a interiorizar los principios matemáticos y computacionales implicados en la seguridad de la información en la era post-cuántica.**

***Index Terms*—Criptografía post-cuántica, gamificación, retículos, LWE, CRYSTALS-Kyber**

Tipo de contribución: *Formación e innovación educativa*

I. Introducción

La revolución de la computación cuántica promete un cambio de paradigma tecnológico al resolver en tan solo cuestión de segundos problemas matemáticos que tomarían miles de años a los ordenadores actuales más potentes. Sin embargo, con estos avances también llega una amenaza significativa a la ciberseguridad. La computación cuántica podría usarse para romper los sistemas de cifrado actuales que protegen nuestros datos bancarios, comunicaciones y secretos gubernamentales. Esto ha implicado el inicio de una carrera contra reloj para desarrollar nuevos métodos criptográficos resistentes a la era cuántica, asegurando un futuro digital seguro.

Desde 2016, el Instituto Nacional de Estándares y Tecnología (*National Institute of Standards and Technology, NIST*) ha dedicado un gran esfuerzo al proceso de selección de algoritmos que sean resistentes a la cuántica. A finales de 2017 se publicaron los algoritmos que habían pasado la primera fase de la convocatoria, un total de 69. Más tarde, en 2019 tras aplicar un cribado exhaustivo quedaron 26. Durante esa ronda se llevaron a cabo pruebas rigurosas para evaluar la seguridad y el rendimiento de cada uno de los algoritmos, así como para identificar posibles vulnerabilidades y mejorar la seguridad de los algoritmos. En la tercera ronda de evaluación en 2018, se seleccionaron 15 candidatos finales para continuar en el proceso de estandarización. Durante esta ronda, se llevaron a cabo pruebas adicionales y se trabajó con la comunidad criptográfica para identificar posibles vulnerabilidades y mejorar la seguridad de los algoritmos preseleccionados. Finalmente, en julio de 2022, el NIST anunció los algoritmos seleccionados como estándares finales [1]. La propuesta de estandarización incluye:

- Cifrado: CRYSTALS-Kyber [2].
- Firma digital: CRYSTALS-Dilithium [3], FALCON [4] y SPHINCS+ [5].

Además, a mediados de 2022, los esquemas de cifrado BIKE, Classic McElice, HQC (Hamming Quasi-Cyclic) y SIKE pasaron a una cuarta ronda [6], para continuar su análisis y evaluación [7].

En cuanto a la firma digital, el siguiente paso del NIST fue diversificar su abanico de esquemas, incluyendo alternativas no basadas en retículos [8]. Esta estrategia responde al principio de criptoagilidad, ya que garantiza la capacidad de los sistemas de seguridad para adaptarse rápidamente a cambios en los algoritmos criptográficos, promoviendo para ello la coexistencia de distintas familias criptográficas como medida preventiva ante futuras vulnerabilidades. En particular, se priorizó la búsqueda de esquemas con firmas cortas y verificación rápida, favoreciendo así su aplicación en contextos con restricciones de recursos o requisitos de alta eficiencia.

Ya en 2023, el NIST redactó los borradores de las Normas Federales de Tratamiento de la Información (*Federal Information Processing Standards, FIPS*) FIPS 203, FIPS 204 y FIPS 205 [9], [10], [11], en las que se define CRYSTALS-Kyber como el algoritmo de encapsulación de claves basado en retículos (*Module Lattice Key Encapsulation Mechanism, ML-KEM*), CRYSTALS-Dilithium como el algoritmo de firma digital basado en retículos (*Module Lattice Digital Signature Algorithm, ML-DSA*) y SPHINCS+ como el algoritmo de firma digital basado en hash. A finales de 2024 se concluyeron los borradores y se publicaron las versiones finales.

Recientemente, en 2025, el NIST publicó el estándar HQC [12], [13], un mecanismo de encapsulación de claves fundamentado en códigos de corrección de errores.

Tras este extenso proceso de evaluación y estandarización, gran parte de la atención de la comunidad criptográfica se ha centrado en el algoritmo CRYSTALS-Kyber y en la estructura matemática que lo sustenta: los retículos. Este esquema ha sido diseñado como una alternativa post-cuántica frente a los sistemas criptográficos asimétricos tradicionales, tales como RSA o ECDSA, cuya seguridad se basa en problemas como la factorización de enteros o el cálculo del logaritmo discreto, respectivamente. A diferencia de estos enfoques clásicos, la seguridad de Kyber se fundamenta en la complejidad computacional de problemas sobre retículos, lo que le confiere resistencia frente a ataques cuánticos. En concreto, define un mecanismo de encapsulación de claves que satisface

la propiedad de seguridad IND-CCA2 (Indistinguishability under Adaptive Chosen Ciphertext Attack), es decir, que garantiza que un atacante no pueda distinguir entre diferentes mensajes cifrados, incluso si se le permite realizar consultas sobre el descifrado de otros mensajes. La robustez de este esquema se basa en el problema de Aprendizaje con Errores en Módulos (*Module Learning With Errors, MLWE*) [14], una generalización del problema de Aprendizaje con Errores (*Learning With Errors, LWE*) a estructuras algebraicas más complejas, como los módulos sobre anillos.

Si bien la estandarización de algoritmos capaces de resistir los ataques basados en computación cuántica ha representado un avance significativo, persiste un desafío igualmente crucial: comprender y hacer comprensibles los fundamentos matemáticos que subyacen en estos nuevos esquemas criptográficos. Su complejidad algebraica y estructural constituye una barrera importante para su introducción en contextos educativos. Hasta el momento, la enseñanza de contenidos vinculados a la criptografía post-cuántica apenas ha sido explorada en el ámbito científico, y existen pocas iniciativas que aborden su integración curricular desde una perspectiva innovadora. Sin embargo, todo indica que, en los próximos años, el abordaje didáctico de estos conceptos será clave en disciplinas como las matemáticas, física e ingeniería. En este sentido, este trabajo surge como respuesta a esta necesidad emergente y se plantea como una continuación de la propuesta presentada en [15], ampliando su alcance hacia el diseño de recursos interactivos que favorezcan la comprensión de conceptos criptográficos avanzados desde etapas formativas iniciales.

Este trabajo se estructura de la siguiente forma. En primer lugar, en la Sección II se presentan las herramientas necesarias para comprender el trabajo. Se expone qué es Genially, se justifica para qué tipo de alumnado está enfocada la actividad y se introducen los conceptos matemáticos que se trabajan en el ejercicio. Estos conceptos se introducen de manera formal ya que son los que debe conocer el docente antes de la actividad. Luego, en la Sección III se desarrolla y explica la actividad, que puede consultarse en [16]. Por último, se dedica una última sección a los trabajos futuros y las conclusiones.

II. Preliminares

Los esquemas criptográficos basados en retículos y en el problema LWE constituyen actualmente la base más extendida en los estándares de cifrado y firma digital resistentes a ataques cuánticos. No obstante, la literatura educativa relacionada con estos conceptos es aún escasa, y los recursos disponibles suelen presentar un nivel de abstracción técnica elevado, lo que dificulta su accesibilidad para estudiantes y docentes sin formación especializada en criptografía. En esta situación, este trabajo propone una serie de actividades interactivas orientadas a facilitar la comprensión de los fundamentos matemáticos subyacentes, mediante representaciones visuales y elementos didácticos interactivos. Dichas actividades han sido diseñadas y publicadas en la plataforma Genially [17], con el objetivo de ofrecer una herramienta dinámica y accesible para el aula.

Herramienta tecnológica

Genially es una plataforma digital desarrollada por una empresa española fundada en 2015 [18], con el propósito de transformar los modos de comunicación visual e impulsar la innovación educativa. Se trata de una herramienta en línea que permite crear contenidos interactivos de manera intuitiva, sin necesidad de conocimientos avanzados en diseño o programación. Su versatilidad abarca la elaboración de presentaciones, infografías, juegos, mapas interactivos, materiales de e-learning y otros recursos digitales, facilitando la generación de experiencias visuales y participativas tanto en contextos educativos como profesionales. Esta accesibilidad tecnológica la convierte en un entorno idóneo para el desarrollo de propuestas didácticas dirigidas a la enseñanza de conceptos matemáticos complejos, como los que subyacen en la criptografía post-cuántica.

Diversos estudios han producido resultados positivos en el aprendizaje mediante el uso de Genially como plataforma educativa. Por ejemplo, en [19], se aplicó una estrategia basada en gamificación con Genially, concluyéndose que el uso de recursos visualmente atractivos y clases más dinámicas mejoró notablemente la percepción y aceptación de las matemáticas por parte del alumnado. En otro trabajo, [20], se exploró la integración de aprendizaje basado en problemas y gamificación con Genially en un curso universitario de Fundamentos de Matemáticas para Ingeniería Industrial, obteniéndose mejoras significativas tanto en las habilidades de resolución de problemas como en el rendimiento académico promedio de los estudiantes.

A la luz de estos antecedentes y de las ventajas que ofrece Genially como entorno de diseño interactivo, se ha optado por utilizar esta plataforma para el desarrollo de una serie de actividades introductorias sobre criptografía post-cuántica, centradas particularmente en el problema de aprendizaje con errores. El objetivo es facilitar la comprensión de los conceptos matemáticos subyacentes mediante experiencias visuales e interactivas que favorezcan el aprendizaje activo desde etapas educativas tempranas.

Justificación curricular

Esta propuesta puede ser implementada a partir del nivel de Bachillerato, específicamente en la asignatura de Matemáticas II, donde el alumnado ya dispone de los conocimientos necesarios —como el trabajo con vectores, productos escalares y espacios euclídeos— para abordar y comprender de forma introductoria problemas del tipo LWE. La incorporación de esta innovación educativa se justifica curricularmente en función de los contenidos y competencias recogidos en la normativa vigente. Para ello, se ha tomado como referencia el marco establecido en el Real Decreto 243/2022, por el que se fijan los aspectos básicos del currículo del Bachillerato en el sistema educativo español [21]. En concreto, esta propuesta se alinea con varios de los elementos curriculares recogidos en dicho Decreto, tanto en lo relativo a las competencias específicas como a los saberes básicos de la asignatura de Matemáticas II. A continuación, se enumeran aquellos que guardan una vinculación más estrecha con los objetivos y contenidos de esta propuesta:

- Competencias específicas.
  - Competencia específica 4: Utilizar el pensamiento computacional de forma eficaz, modificando, creando y generalizando algoritmos que resuelvan problemas mediante el uso de las matemáticas, para

modelizar y resolver situaciones de la vida cotidiana y del ámbito de la ciencia y la tecnología.
- Competencia específica 5: Establecer, investigar y utilizar conexiones entre las diferentes ideas matemáticas estableciendo vínculos entre conceptos, procedimientos, argumentos y modelos para dar significado y estructurar el aprendizaje matemático.
- Competencia específica 6: Descubrir los vínculos de las matemáticas con otras áreas de conocimiento y profundizar en sus conexiones, interrelacionando conceptos y procedimientos, para modelizar, resolver problemas y desarrollar la capacidad crítica, creativa e innovadora en situaciones diversas.
- Competencia específica 7: Representar conceptos, procedimientos e información matemáticos seleccionando diferentes tecnologías, para visualizar ideas y estructurar razonamientos matemáticos.

- Saberes básicos.
  - Sentido algebraico.
    - 2. Modelo matemático. Sistemas de ecuaciones: modelización de situaciones en diversos contextos.
    - 3. Igualdad y desigualdad. Formas equivalentes de expresiones algebraicas en la resolución de sistemas de ecuaciones e inecuaciones, mediante cálculo mental, algoritmos de lápiz y papel, y con herramientas digitales. Resolución de sistemas de ecuaciones en diferentes contextos.
    - 6. Pensamiento computacional. Análisis algorítmico de las propiedades de las operaciones con matrices, los determinantes y la resolución de sistemas de ecuaciones lineales.
  - Sentido estocástico.
    - 2. Distribuciones de probabilidad. Modelización de fenómenos estocásticos mediante las distribuciones de probabilidad binomial y normal. Cálculo de probabilidades asociadas mediante herramientas tecnológicas.

Además, este recurso se adscribe al ámbito del conocimiento 'Matemáticas y estadística' según el Anexo I del RD 822/2021 [22].

Por otra parte, resulta relevante señalar que una de las principales motivaciones para el desarrollo de este trabajo ha sido el último Informe PISA 2022 [23], en el que España ha obtenido su puntuación más baja desde el inicio de esta evaluación internacional. En particular, el alumnado español de 4º de Educación Secundaria Obligatoria ha experimentado una disminución de 8 puntos en competencia matemática respecto a la edición anterior. Esta tendencia descendente provoca una preocupación creciente en los docentes por el rendimiento matemático en etapas obligatorias y post-obligatorias.

A la luz de esos resultados, se hace evidente la necesidad urgente de desarrollar materiales didácticos innovadores y eficaces que permitan abordar las deficiencias detectadas en el aprendizaje de las matemáticas. En este contexto, el presente trabajo busca contribuir al fortalecimiento de las competencias numéricas y conceptuales del alumnado, integrando contenidos de actualidad como la criptografía post-cuántica, y vinculándolos con el currículo de matemáticas de Bachillerato a través de propuestas pedagógicas activas, visuales e interactivas.

Conceptos matemáticos

La actividad desarrollada en [24] persigue un doble objetivo didáctico. Por un lado, se propone introducir la criptografía post-cuántica —y en particular el problema LWE— a partir de los principios matemáticos que se enseñan en Educación Secundaria y Bachillerato. Por otro lado, se plantea utilizar el contexto de la criptografía post-cuántica como medio para reforzar la comprensión de diversos conceptos matemáticos curriculares, promoviendo así una conexión significativa entre teoría y aplicación.

Los objetivos establecidos en esta actividad didáctica poseen una gran relevancia en el contexto educativo contemporáneo, no solo por su gran valor formativo, sino también porque están alineados con algunas de las corrientes pedagógicas más innovadoras. Primero, el hecho de que se introduzcan conceptos de criptografía post-cuántica como el LWE, desde la secundaria y la enseñanza de bachillerato, se debe a la posibilidad de acercar a los estudiantes a un campo de investigación actual, cuyas aplicaciones reales en la ciberseguridad y los desafíos tecnológicos del mundo actual. En este sentido, es más fácil que los jóvenes se interesen en las matemáticas si se les explicaran desde una perspectiva utilitaria y contemporáneamente aplicada, de vital importancia para incentivar posibles vocaciones científicas y tecnológicas.

Por otro lado, esta propuesta es congruente con el enfoque STEM (*Science, Technology, Engineering and Mathematics*). Desde una perspectiva interdisciplinaria reúne conocimiento vital en áreas de aprendizaje siguiendo un esquema de educación por competencias.

Además, el hecho de que se aborde el tema de la criptografía post-cuántica con la intención de exponer conceptos del currículo del nivel de secundaria como álgebra lineal básica, operaciones con vectores o tratar el error, conforma un aprendizaje significativo [25] que se da cuando los nuevos conocimientos se integran con estructuras previas del pensamiento del estudiante.

Finalmente, se trata de una propuesta de activación por competencias, puesto que responde a una contingencia específica que requiere de investigación, exploración y aplicación por parte de los estudiantes. Por lo tanto, se fomenta un papel activo en procesos de autoaprendizaje, fomentando la autonomía del estudiante.

En resumen, estos objetivos trazados no solo contribuyen al currículo, sino que también lo completan con la contextualización y la metodología propuesta.

Entre los contenidos matemáticos que se trabajan en la propuesta destacan los siguientes:

- Producto escalar de dos vectores
- Estructura de retículos
- Resolución de sistemas de ecuaciones lineales
- Manipulación algebraica de polinomios

A continuación, se presentan de forma formal y progresiva los conceptos matemáticos implicados en la actividad y que debe conocer el docente, con el fin de facilitar su contextualización didáctica y su posterior aplicación al estudio del problema LWE.

Definición 1 (Retículo): Sean V un espacio vectorial sobre un cuerpo K, $B = \{v_1, v_2, \ldots, v_n\}$ una base de un subespacio vectorial de V. Entonces el retículo $\mathcal{L} \subset V$ generado por la base B es el conjunto:

$$\mathcal{L}(B) = \{ \sum_{i=1}^{n} a_i \cdot v_i : a_i \in \mathbb{Z} \} \tag{1}$$

Por tanto, un retículo es el conjunto de puntos generado por todas las combinaciones lineales enteras de los vectores de una base en el espacio vectorial en el que se define.

Considérese un vector secreto de coeficientes enteros $s = (s_1, s_2, \ldots, s_n) \in \mathbb{Z}^n$ y un sistema lineal de m ecuaciones con coeficientes conocidos cuya solución es s (con $m \geq n$), esto se puede expresar como:

$$\begin{cases} b_1 = a_{1,1}s_1 + a_{1,2}s_2 + \ldots + a_{1,n}s_n \\ b_2 = a_{2,1}s_1 + a_{2,2}s_2 + \ldots + a_{2,n}s_n \\ \vdots \\ b_m = a_{m,1}s_1 + a_{m,2}s_2 + \ldots + a_{m,n}s_n \end{cases} \tag{2}$$

Resolver este sistema es relativamente sencillo aplicando cualquier método adecuado para esta tarea, como la *eliminación gaussiana*. Sin embargo, la situación cambia considerablemente si se introduce una ligera modificación. En concreto, se puede añadir un pequeño error en cada ecuación, lo que da lugar al siguiente sistema:

$$\begin{cases} b_1 = a_{1,1}s_1 + a_{1,2}s_2 + \ldots + a_{1,n}s_n + e_1 \\ b_2 = a_{2,1}s_1 + a_{2,2}s_2 + \ldots + a_{2,n}s_n + e_2 \\ \vdots \\ b_m = a_{m,1}s_1 + a_{m,2}s_2 + \ldots + a_{m,n}s_n + e_m \end{cases} \tag{3}$$

Si bien el sistema resultante deja de ser exacto, la introducción de pequeños errores impide aplicar los métodos clásicos de álgebra lineal para obtener una solución directa. No obstante, si los errores son suficientemente pequeños, el sistema puede ser aproximado, aunque su resolución se convierte en un problema computacionalmente complejo. En este contexto, se puede expresar el sistema modificado de la siguiente manera:

$$\begin{cases} b_1 \approx a_{1,1}s_1 + a_{1,2}s_2 + \ldots + a_{1,n}s_n \\ b_2 \approx a_{2,1}s_1 + a_{2,2}s_2 + \ldots + a_{2,n}s_n \\ \vdots \\ b_m \approx a_{m,1}s_1 + a_{m,2}s_2 + \ldots + a_{m,n}s_n \end{cases} \tag{4}$$

Resolver este sistema ya no es una tarea sencilla, ya que al aplicar técnicas como la reducción por filas, los errores introducidos en las ecuaciones pueden acumularse progresivamente e incluso de forma exponencial. Esta acumulación puede provocar que la solución obtenida se aleje considerablemente del valor original, lo que dificulta o incluso imposibilita la recuperación exacta de las variables desconocidas.

Formalización del problema LWE: Sean $n \in \mathbb{N}$ y $q \in \mathbb{Z}$, cuyo tamaño es similar a n. Considérese además un conjunto de m vectores $b_1, b_2, \ldots, b_m \in \mathbb{Z}_q^n$.

Fijada una distribución de probabilidad $\mathcal{X}$ sobre $\mathbb{Z}_q$, esta permite seleccionar un término de error de manera controlada. La selección de un error e de acuerdo con esta distribución se denota como $e \leftarrow \mathcal{X}$. Estos parámetros sirven para definir la llamada *distribución LWE* sobre $\mathbb{Z}_q^n \times \mathbb{Z}_q$ (Definición 2).

Definición 2 (Distribución LWE): Sean $n, q \in \mathbb{N}$, $s \in \mathbb{Z}_q^n$ un vector secreto fijado de antemano, y $\mathcal{X}$ una distribución de probabilidad sobre $\mathbb{Z}_q$. Se define la *distribución LWE módulo q asociada a s*, denotada $\mathcal{A}_{s,\mathcal{X}}$, como la distribución sobre pares $(\mathbf{A}, b) \in \mathbb{Z}_q^n \times \mathbb{Z}_q$ obtenida mediante el siguiente proceso:

1. Se elige un vector $\mathbf{A} \in \mathbb{Z}_q^n$ uniformemente al azar.
2. Se selecciona un error $e \leftarrow \mathcal{X}$.
3. Se calcula
$$b \;=\; \langle \mathbf{A},\, s \rangle + e \mod q.$$
4. Se devuelve el par $(\mathbf{A}, b)$.

De este modo, la distribución genera pares $\langle A, b \rangle \in \mathbb{Z}_q^n \times \mathbb{Z}_q$. A partir de esta definición, resulta natural plantear una versión del problema LWE en términos de un problema de búsqueda. Concretamente, este nuevo problema se puede formular del siguiente modo:

Definición 3 (Problema de búsqueda LWE): Sean $n, m, q \in \mathbb{N}$, y sea s un vector escogido uniformemente al azar en $\mathbb{Z}_q^n$. Consideremos además la distribución $\mathcal{A}_{s,\mathcal{X}}$ definida anteriormente. El problema consiste en determinar el vector secreto s a partir de m muestras independientes extraídas de la distribución $\mathcal{A}_{s,\mathcal{X}}$.

Según [26], dados $n, m, q \in \mathbb{N}$, el problema de aprendizaje con errores puede plantearse en términos algebraicos como la tarea de encontrar un vector secreto $s \in \mathbb{Z}_q^n$ que satisfaga la ecuación $As + e = b \mod q$, donde $A \in \mathbb{Z}_q^{m \times n}$ es una matriz conocida, mientras que $e \in \mathbb{Z}_q^m$ es un vector de error obtenido a partir de una distribución de probabilidad $\mathcal{X}^m$ sobre $\mathbb{Z}_q^m$, y $b \in \mathbb{Z}_q^m$ representa el vector resultante.

En particular, se considera el anillo $\mathcal{R}_q = \mathbb{Z}_q[x]/\langle x^n + 1 \rangle$, con $n = 2^k$, de modo que el polinomio $(x^n + 1)$ corresponde al polinomio ciclotómico de orden 2^{k+1}. En este contexto, el par (A, b) queda definido mediante elementos del anillo $\mathcal{R}_q$ de la siguiente forma: $A \in \mathcal{R}_q$ y $b = As + e \mod q$, donde tanto el vector secreto s como el error e pertenecen al anillo $\mathcal{R}_q$, siendo e un término extraído según una distribución de probabilidad $\mathcal{X}$ sobre $\mathcal{R}_q$. Esta versión generalizada del problema LWE se denomina *Module-LWE*.

Aplicaciones criptográficas del problema LWE: Numerosas construcciones criptográficas actuales fundamentan su seguridad en el problema estándar de aprendizaje con errores [14], [27]. Gran parte de estas aplicaciones pueden optimizarse al adaptarlas al problema LWE sobre anillos (*Ring Learning With Errors, RLWE*). Este proceso de adaptación suele ser relativamente sencillo, aunque en algunos casos requiere el uso de herramientas técnicas adicionales para alcanzar resultados óptimos en términos de eficiencia y seguridad práctica.

A continuación, se presenta un ejemplo de la aplicación del problema LWE en CRYSTALS-Kyber, el cual ha sido desarrollado específicamente como parte de la actividad propuesta.

El algoritmo de generación de claves comienza seleccionando uniformemente al azar un elemento $A \in \mathcal{R}_q$ y dos elementos adicionales, $s, e \in \mathcal{R}_q$, extraídos de una distribución de errores apropiada con valores pequeños. La clave

secreta queda definida como s, mientras que la clave pública corresponde al par (A, t), con $t = A \cdot s + e \mod q$.

Para cifrar un mensaje m de n bits, representado por $m = (b_0, b_1, \ldots, b_{n-1})$, donde cada $b_i \in \{0, 1\}$, el algoritmo de cifrado selecciona tres elementos adicionales $r, e_1, e_2 \in \mathcal{R}_q$ con valores pequeños, extraídos de la distribución de errores mencionada. A partir de estos valores, se obtiene el par cifrado (u, v) mediante el siguiente cálculo:

$$u = A \cdot r + e_1 \mod q, \quad v = t \cdot r + e_2 + \left\lceil \frac{q}{2} \right\rceil \cdot m \mod q. \quad (5)$$

Finalmente, para descifrar el mensaje cifrado (u, v), se computa:

$$v - u \cdot s = (r \cdot e - s \cdot e_1 + e_2) + \left\lceil \frac{q}{2} \right\rceil \cdot m \mod q. \quad (6)$$

Para una elección adecuada de los parámetros, los coeficientes del término $(r \cdot e - s \cdot e_1 + e_2)$ presentan magnitudes inferiores a $\frac{q}{4}$. Bajo estas condiciones, cada bit del mensaje original m puede recuperarse mediante un procedimiento de redondeo aplicado a los coeficientes del término $v - u \cdot s$. En concreto, cada coeficiente se redondea al valor más cercano entre 0 y $\left\lceil \frac{q}{2} \right\rceil$. Aquellos coeficientes redondeados a 0 se interpretan como bits 0, mientras que los redondeados a $\left\lceil \frac{q}{2} \right\rceil$ se interpretan como bits 1, permitiendo así recuperar de manera precisa el mensaje original m.

Correctitud de LWE: Obsérvese que, si no existiera el error introducido en las muestras LWE, la expresión $t - A \cdot s$ tomaría únicamente los valores 0 o $\left\lfloor \frac{q}{2} \right\rfloor$, dependiendo del bit cifrado, lo que garantizaría un descifrado siempre correcto. Por lo tanto, [28] indica que solamente se produce un error en el proceso de descifrado cuando la suma de los términos de error sobre todos los coeficientes supera el valor $\frac{q}{4}$.

Dado que se suman como máximo n términos de error independientes, cada uno siguiendo una distribución aproximadamente normal con desviación típica $\alpha \cdot q$, la desviación típica de dicha suma es a lo sumo $\sqrt{m} \cdot \alpha \cdot q \ll \frac{q}{\log(n)}$ acorde a [27]. Un análisis probabilístico estándar demuestra que la probabilidad de que esta variable normal exceda el umbral $\frac{q}{4}$ es despreciable.

III. Propuesta de actividad

La actividad desarrollada en [16] se plantea con dos propósitos: emplear los principios matemáticos enseñados en Secundaria y Bachillerato para introducir la criptografía post-cuántica y el problema de aprendizaje con errores a los estudiantes, y utilizar la criptografía post-cuántica como herramienta para mejorar la comprensión de diversos conceptos matemáticos abordados en el aula.

Al presentar dicho recurso no es necesario que se impartan los contenidos matemáticos descritos en este trabajo. En cambio, la actividad puede servir como una manera alternativa de introducir esos conceptos o como un complemento motivador para reforzar los temas ya abordados en el aula.

III-A. Descripción de la actividad

La actividad desarrollada en Genially titulada "Ataque cuántico a la base post-cuántica", tiene por objetivo ofrecer desde la gamificación una introducción a la criptografía post-cuántica práctica, reforzando los conocimientos previos vistos sobre retículos en [24] y en este caso una ampliación del problema de aprendizaje con errores.

La actividad comienza con las pantallas incluidas en Fig. 1 en las que se proporciona un estado de la misión al jugador. Para esta actividad se sitúa al jugador en una base ficticia que está siendo atacada por un ordenador cuántico. Este ordenador cuántico pretende vulnerar la seguridad de la base y descifrar todos los datos sensibles. No obstante, un grupo de especialistas formado por matemáticos e ingenieros recomienda una posible solución a la grave situación que se está sufriendo: la criptografía post-cuántica.



Figura 1: Descripción de la misión

Además, se presenta un mapa del juego que representa distintas zonas de una base, las cuales se desbloquean de forma progresiva conforme el jugador supera los desafíos. Para iniciar la resolución de los enigmas planteados, se

propone una situación inicial de estudio y reflexión. A través de un menú interactivo, el jugador puede seleccionar la información necesaria para avanzar en el juego y tomar decisiones fundamentadas (véase Fig. 2).



Figura 2: Introducción al concepto de retículo

La dinámica del juego es constante y repite el patrón *Zona de la base nueva - Escenario de la base - Menú interactivo didáctico - Preguntas*. Hay cinco zonas a lo largo de toda la base y cada zona se dedica a un elemento de la criptografía post-cuántica. A continuación se describen dichas zonas:

- Zona 1. Sección cimientos. En esta parte se trabaja y estudia la matemática que sustenta a los actuales estándares de criptografía post-cuántica: los retículos.
- Zona 2. Sección NIST (Fig. 3). En esta parte el estudiante conoce y explora el desarrollo del proceso de estandarización de la criptografía post-cuántica llevado a cabo por el NIST. Para ello se proporcionan recursos de investigación sobre este proceso. En la parte evaluativa se pregunta sobre CRYSTALS-Kyber, CRYSTALS-Dilithium y HQC.
- Zona 3. Sección instalaciones. En esta parte se introduce



Figura 3: Introducción de los estándares

al estudiante por primera vez en el problema de aprendizaje con errores. Se proporcionan recursos interactivos basados en plataformas como GeoGebra [29] con el objetivo de explicar dinámicamente el efecto de añadir ruido controlado a un sistema de ecuaciones. Para la parte evaluativa se pregunta al jugador sobre la estructura matemática que conforman el problema de aprendizaje con errores y sobre la formulación de dicho problema (Fig. 4).
- Zona 4. Sección codificación. En esta parte el jugador es consciente de que debe asegurar la base y hacer una transición a la post-cuántica. Para ello, debe conocer las especificaciones de CRYSTALS-Kyber y CRYSTALS-Dilithium. Adicionalmente se trabaja el concepto de KEM, así como sus diferentes parámetros para los distintos niveles de seguridad requeridos.
- Zona 5. Sección datos sensibles. En esta misión final se utiliza la calculadora desarrollada en [30] para poder llevar a cabo los cálculos. De este modo, el jugador demuestra que reunir los conocimientos aprendidos durante el juego, reconociendo la estructura del problema

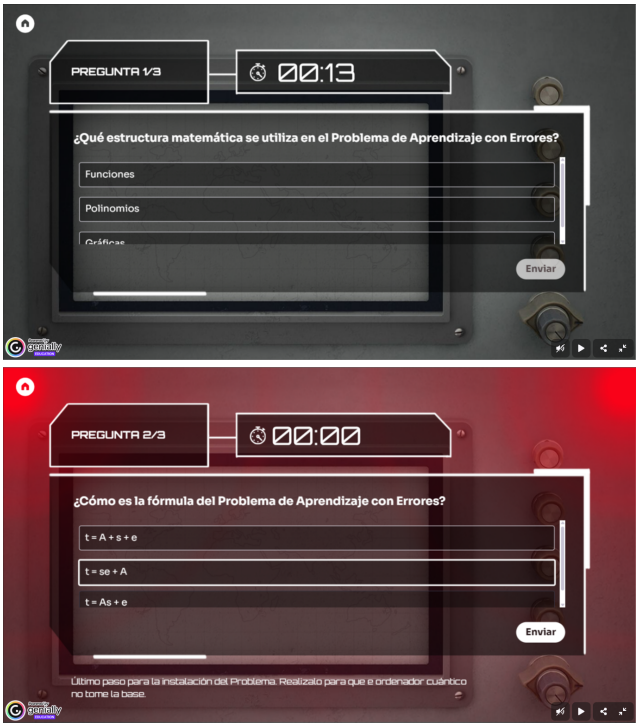


Figura 4: Evaluación del proceso de aprendizaje

de aprendizaje con errores, conociendo cómo se escala un polinomio y operando con polinomios (Fig. 5). Con la finalidad de evaluar el proceso de aprendizaje, se incluyen preguntas relacionadas con los resultados proporcionados por la calculadora.



Figura 5: Visualización de la pantalla final

Una vez el jugador haya completado todas las tareas, habrá asegurado la base de la amenaza cuántica.

III-B. Evaluación de la actividad

Al finalizar la misión gamificada, se propone realizar un juego de Kahoot con preguntas de respuesta múltiple sobre los conocimientos impartidos en la actividad. Esta evaluación ha sido diseñada para reforzar los aprendizajes del juego y evaluar si han adquirido los conocimientos a lo largo de la sesión. Para esta evaluación se proponen las siguientes preguntas:

1. ¿Cuál es el objetivo principal de la criptografía post-cuántica?
   - Proteger la información frente a adversarios con ordenadores cuánticos
   - Romper más rápidamente los esquemas RSA y ECC
   - Reducir el tamaño de las claves de cifrado simétrico
   - Acelerar la generación de números aleatorios criptográficos
2. ¿Qué es el NIST?
   - El Instituto Nacional de Estándares y Tecnología
   - El Instituto Nacional de Seguridad y Telecomunicaciones
   - El Instituto Nacional de Sistemas y Tecnología
   - El Instituto Nacional de Estadísticas y Telecomunicaciones
3. ¿Para qué sirve el algoritmo CRYSTALS-Kyber?
   - Es un algoritmo para cifrar.
   - Es un algoritmo de firma digital.
   - Es un algoritmo para multiplicar polinomios eficientemente.
   - Es un algoritmo que no resiste a los ordenadores cuánticos.
4. ¿Cómo se formula el problema de *Learning with errors*?
   - $A + s + e$
   - $A + e \cdot s$
   - $A \cdot s + e$
   - $A \cdot e + s$
5. ¿Qué representa el vector s en LWE?
   - Representa un vector de aproximación.
   - Representa un vector de error.
   - Representa la matriz de coeficientes de seguridad.
   - Representa un vector secreto.
6. ¿Cuáles de los siguientes algoritmos es una firma digital?
   - CRYSTALS-Kyber
   - Classic McEliece
   - CRYSTALS-Dilithium
   - Cifrado César
7. Cuando se aplica el *Learning with errors* al esquema de CRYSTALS-Kyber es necesario:
   - Reescalar el mensaje en binario que se quiere cifrar a $\mod \lfloor \frac{q}{2} \rfloor$.
   - Ajustar el nivel de ruido antes de cifrar.
   - Convertir el mensaje en un polinomio sobre $\mathbb{Z}_q$.
   - Generar la clave pública del receptor antes de añadir el error.
8. Cuando se va a descifrar el mensaje en CRYSTALS-Kyber se obtiene:
   - Un mensaje ruidoso que después hay que aproximar.
   - El mensaje original en claro sin ruido.
   - La clave privada del receptor.
   - El vector de error aislado.

En definitiva, esta evaluación, basada en dinámicas lúdicas (Kahoot) y tareas prácticas, no solo comprueba de forma continua la asimilación de conceptos clave y el desarrollo de competencias, sino que también proporciona información

inmediata para ajustar la enseñanza y reforzar los contenidos impartidos.

IV. Conclusiones

En este trabajo se ha presentado un enfoque innovador para la enseñanza de la criptografía post-cuántica mediante el uso de la plataforma Genially. A lo largo de la actividad propuesta, se abordan conceptos clave como los retículos, el proceso de estandarización del NIST, las bases matemáticas del problema LWE, así como las especificaciones de los esquemas estándares basados en CRYSTALS-Kyber y CRYSTALS-Dilithium.

Como cierre, se ha desarrollado una aplicación práctica y accesible de CRYSTALS-Kyber mediante una calculadora interactiva, lo que permite una comprensión más intuitiva del proceso de cifrado. Con respecto a la revisión realizada para este trabajo, no se han encontrado propuestas similares con un enfoque didáctico interactivo en este campo.

Como trabajo en desarrollo, actualmente se está planificando llevar a cabo talleres con estudiantes para evaluar la eficacia y utilidad del recurso propuesto, además de introducir mejoras y adaptar la calculadora a una interfaz gráfica más dinámica e intuitiva.

Agradecimientos

Esta investigación ha sido posible gracias a los proyectos PID2022-138933OB-I00 ATQUE, 2023DIG28 IACTA y C064/23 SCITALA ULL-INCIBE, y la Cátedra de Ciberseguridad C065/23 ULL-INCIBE, financiados por MCIN/AEI/10.13039/501100011033, Unión Europea Next-GenerationEU/PRTR y Fundación CajaCanarias-La Caixa.

Referencias

[1] "NIST announces first four quantum-resistant cryptographic algorithms." [Online]. Available: https://www.nist.gov/news-events/news/2022/07/nist-announces-first-four-quantum-resistant-cryptographic-algorithms

[2] R. Avanzi, J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. Schanck, P. Schwabe, G. Seiler, and D. Stehlé, "CRYSTALS-Kyber algorithm specifications and supporting documentation," *NIST PQC Round*, vol. 2, no. 4, pp. 1–43, 2021.

[3] L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, P. Schwabe, G. Seiler, and D. Stehlé, "CRYSTALS-Dilithium: Algorithm specifications and supporting documentation (version 3.1)," *NIST Post-Quantum Cryptography Standardization Round*, vol. 3, pp. 1–38, 2021.

[4] P.-A. Fouque, J. Hoffstein, P. Kirchner, V. Lyubashevsky, T. Pornin, T. Prest, T. Ricosset, G. Seiler, W. Whyte, and Z. Zhang, "Falcon: Fast-fourier lattice-based compact signatures over ntru," *Submission to the NIST's post-quantum cryptography standardization process*, vol. 36, no. 5, pp. 1–75, 2018.

[5] J.-P. Aumasson, D. J. Bernstein, W. Beullens, C. Dobraunig, M. Eichlseder, S. Fluhrer, S.-L. Gazdag, A. Hülsing, P. Kampanakis, S. Kölbl, T. Lange, M. M. Lauridsen, F. Mendel, R. Niederhagen, C. Rechberger, J. Rijneveld, P. Schwabe, and B. Westerbaan, "SPHINCS+ Specification: Submission to the NIST Post-Quantum Cryptography Standardization Project (Round 3.1)," https://sphincs.org/data/sphincs+-r3.1-specification.pdf, 2021, disponible en: https://sphincs.org/data/sphincs+-r3.1-specification.pdf.

[6] National Institute of Standards and Technology, "Pqc standardization process: Announcing four candidates to be standardized, plus fourth round candidate," Jul. 2022.

[7] National Institute of Standards and Technology (NIST), "Post-quantum cryptography, round 4 submissions," https://csrc.nist.gov/projects/post-quantum-cryptography/round-4-submissions, Feb 2025, Último acceso: Marzo de 2025.

[8] National Institute of Standards and Technology, "Post-quantum cryptography: Additional digital signature schemes," https://csrc.nist.gov/projects/pqc-dig-sig/round-2-additional-signatures, Feb 2025, Último acceso: Marzo de 2025.

[9] National Institute of Standards and Technology, "Module-lattice-based key-encapsulation mechanism standard," Gaithersburg, MD 20899-8900, 2024.

[10] National Institute of Standards and Technology, "Module lattice based digital signature standard," Gaithersburg, MD 20899-8900, 2024.

[11] National Institute of Standards and Technology, "Stateless hash-based digital signature standard," https://doi.org/10.6028/NIST.FIPS.205.ipd, Washington, D.C., 2023, federal Information Processing Standards Publication (FIPS) NIST FIPS 205 ipd.

[12] National Institute of Standards and Technology, "Nist selects hqc as the fifth algorithm for post-quantum encryption," https://www.nist.gov/news-events/news/2025/03/nist-selects-hqc-fifth-algorithm-post-quantum-encryption, Mar 2025, Último acceso: Marzo de 2025.

[13] C. Aguilar-Melchor, N. Aragon, S. Bettaieb, L. Bidoux, O. Blazy, J. Bos, J.-C. Deneuville, A. Dion, P. Gaborit, J. Lacan, E. Persichetti, J.-M. Robert, P. Véron, and G. Zémor, "HQC: Hamming Quasi-Cyclic (cuarta ronda, versión actualizada 19 de febrero de 2025)," https://pqc-hqc.org/doc/hqc-specification_2025-02-19.pdf, Feb 2025, disponible en: https://pqc-hqc.org/doc/hqc-specification_2025-02-19.pdf.

[14] A. Langlois and D. Stehlé, "Worst-case to average-case reductions for module lattices," *Designs, Codes and Cryptography*, vol. 75, pp. 565–599, 2015. [Online]. Available: https://doi.org/10.1007/s10623-014-9938-4

[15] E. Perez-Ramos, P. Caballero-Gil, and H. Reboso-Morales, "Wip: Using geogebra to learn the basics of post-quantum cryptography," in *IEEE Frontiers in Education Conference, FIE 2024, Washington DC, USA, October 13-16, 2024*. IEEE, 2024.

[16] E. Pérez-Ramos, C. Hernández-Goya, and P. Caballero-Gil, "Ataque cuántico a la base post-cuántica." https://view.genially.com/67d80511d9ee7dc76780d0b1/interactive-content-teaching-post-quantum-cryptography, 2025.

[17] J. Rubio, L. García, and C. Roldán, "Genially," https://genially.com/es/, 2025, Último acceso: marzo de 2025.

[18] Diario Córdoba, "Nace en córdoba genial.ly, una herramienta que pretende revolucionar la comunicación y la educación," https://www.diariocordoba.com/cordoba-ciudad/2015/06/29/nace-cordoba-genial-ly-herramienta-37026613.html, jun 2015, Último acceso: marzo de 2025.

[19] C. Corte-Rojas and L. Enciso, "Genially a tool for learning mathematics," in *2023 XIII International Conference on Virtual Campus (JICV)*. IEEE, 2023.

[20] E. Boom-Cárcamo, L. Buelvas-Gutiérrez, L. Acosta-Oñate, and D. Boom-Cárcamo, "Gamification and problem-based learning (pbl): Development of creativity in the teaching-learning process of mathematics in university students," *Thinking Skills and Creativity*, vol. 53, 2024.

[21] Gobierno de España. Ministerio de Educación y Formación Profesional y Deportes, "Real decreto 217/2022, de 29 de marzo, por el que se establece la ordenación y las enseñanzas mínimas de la educación secundaria obligatoria," https://www.boe.es, 2022.

[22] Ministerio de Universidades, Gobierno de España, "Real Decreto 822/2021, de 28 de septiembre, por el que se establece la organización de las enseñanzas universitarias y del procedimiento de aseguramiento de su calidad." https://www.boe.es/buscar/act.php?id=BOE-A-2021-15781, Sep. 2021.

[23] OECD, "PISA 2022 Results (Volume IV) - Factsheets: Spain," https://www.oecd.org, 2024.

[24] E. Perez Ramos, "A brief introduction to lattice theory," https://www.geogebra.org/m/js4x7wfj, 2024.

[25] T. Bryce and E. Blown, "Ausubel's meaningful learning re-visited," *Current Psychology*, pp. 1–20, 2023.

[26] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," in *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC)*. ACM, 2005, pp. 84–93.

[27] V. Lyubashevsky, C. Peikert, and O. Regev, "On ideal lattices and learning with errors over rings," *J. ACM*, vol. 60, no. 6, pp. 43:1–43:35, 2013. [Online]. Available: https://doi.org/10.1145/2535925

[28] O. Regev, "The learning with errors problem (invited survey)," in *Proceedings of the 2010 IEEE 25th Annual Conference on Computational Complexity*, ser. CCC '10. USA: IEEE Computer Society, 2010, p. 191–204. [Online]. Available: https://doi.org/10.1109/CCC.2010.26

[29] GeoGebra Team, "GeoGebra - Mathematics Learning Platform," https://www.geogebra.org/, 2024.

[30] E. Pérez-Ramos and P. Hernández-Goya, C.and Caballero-Gil, "Calculadora crystals-kyber," https://colab.research.google.com/drive/1zYjVb56cFIrgrvMCDV3kKsSw8ntjhb0X?usp=sharing, 2025.

Programa científico: pósters

Sesión Pósters I: Detección de amenazas y ciberataques emergentes

A Review of "Exploring Shifting Patterns in Recent IoT Malware"

J. Carrillo-Mondéjar*, G. Suarez-Tangil†, Andrei Costin§ and Ricardo J. Rodríguez‡
*‡Universidad de Zaragoza, §University of Jyväskylä, †IMDEA Networks Institute
{*jcarrillo, ‡rjrodriguez}@unizar.es, †guillermo.suarez-tangil@imdea.org, §ancostin@jyu.fi

***Abstract*—The rise of malware targeting interconnected infrastructures has surged in recent years, driven largely by the widespread presence of vulnerable legacy IoT devices and inadequately secured networks. Despite the strong interest attackers have in targeting this infrastructure, a significant gap remains in understanding how the landscape has recently evolved. Addressing this knowledge gap is essential to thwarting the proliferation of massive botnets, thereby safeguarding end-users and preventing disruptions in critical infrastructures. This work offers a contemporary analysis of Linux-based malware, specifically tailored to IoT malware operating in 2021-2023. Using automated techniques involving both static and dynamic analysis, we classify malware into related threats. By scrutinizing the most recent dataset of Linux-based malware and comparing it to previous studies, we unveil distinctive insights into emerging trends, offering an unparalleled understanding of the evolving landscape. Although Mirai and Gafgyt remain the most prominent families and present a large number of variants, our results show that (i) there is an increase in the sophistication of malware, (ii) malware authors are adding new exploits to their arsenal, and (iii) malware families that originally attacked Windows systems have been adapted to attack Linux-based devices.**

***Index Terms*—Static Analysis; Dynamic Analysis; Malware IoT; Malware Evolution; Malware lineage**

Tipo de contribución: *Investigación ya publicada* en "Exploring Shifting Patterns in Recent IoT Malware" En Proceedings of the European Conference on Cyber Warfare and Security. Academic Conferences International Ltd, 2024 [1].

I. EXTENDED ABSTRACT

The battle against malware has been going on for more than 30 years in the security community. Historically, Windows was the primary target because of its widespread use. However, a significant shift has occurred with the rise of Linux-based operating systems, driven by the growing popularity and the large number of Internet of Things (IoT) devices that use Linux as the underlying operating system.

The global expansion of IoT, which was predicted to reach 50 billion devices by 2020 according to Cisco's [2] and to reach 100 billion by 2025 according to Huawei [3], brings various benefits in various sectors, including industry, transportation, smart cities, health-care, and daily life. The success and attractive features of IoT have revolutionized the technology, but have also raised serious concerns about the security and privacy of user data.

IoT comprises a multitude of interconnected devices within a network, using various communication protocols, such as MQTT, CoAP, and UPnP, among others. None of these systems are inherently designed with security in mind; rather, they prioritize cost-effectiveness and innovation, often at the expense of security and privacy [4]. Additionally, the devices themselves have intrinsic limitations, including low storage and computing capacities, along with challenges such as manufacturers' reluctance to provide software/firmware updates [5] or the complexity involved in installing updates, even for competent users.

Despite innovative features, many IoT devices exhibit vulnerabilities, ranging from default passwords to outdated software with known security issues [6], [7], which are subsequently abused by increasingly sophisticated IoT malware and botnet families [8]. For instance, Mirai infected more than 600,000 devices in 2016 using default passwords from IoT devices such as IP cameras or routers, causing one of the most severe denial-of-service attacks on the Internet [9].

This highlights security challenges within networks of IoT systems, where keeping firmware/software up to date represents a significant hurdle [10]. Cybercriminals exploit known vulnerabilities to scan networks and launch attacks aimed at taking control of devices [11]. Notably, the number of malware attacks targeting interconnected infrastructures has increased significantly in recent years [12], [13]. This is mainly due to the widespread presence of vulnerable IoT devices and poorly secured networks [6]. However, despite attackers' interest in attacking this type of infrastructure, there is still a significant gap in understanding how the landscape has evolved. It is crucial to address this knowledge gap to prevent the proliferation of massive botnets, protect end-users, and prevent disruptions in critical infrastructures.

In this regard, by examining the latest Linux-based malware dataset and comparing it with previous studies, we unveil distinctive insights into emerging trends, offering unparalleled insight into the evolving landscape. We show that understanding the evolutionary dynamics of malware is a crucial step in addressing the challenges posed by new computing platforms, such as those in the IoT, online services, and critical infrastructures. By analyzing the latest IoT malware threats through the lens of a novel — and updated — dataset of over 45K malware samples we shed light on several threats unknown to the community. We set out to conduct our study using well-established methods in the area of IoT malware analysis [14], [8], [15], [16], [17], [13]. In particular, we analyze malware samples to extract data and identify patterns.

Then we use the results of static analysis (e.g., architecture, endianness) to determine which virtual machine is best suited to run the malware sample and understand how the malware interacts with the system. At the end of this stage we obtain execution traces of the samples that were executed successfully. These execution traces consist of the sequence of system calls made by the malware sample, as well as each of the arguments to that system call. We use the execution traces we collect to connect samples that have a similarity greater than 80%. To determine this similarity, we extract N-grams

of size 4 from the set of executed system calls and use the Jaccard index to calculate the similarity between each pair of samples of a particular architecture. We combined the data obtained in both stages to identify the main techniques used by the malware along the following dimensions: obfuscation, exploits used, process interaction, evasion mechanisms, and profiling behaviors. Finally, we generate similarity graphs for each of the architectures and examine each of the unknown clusters in detail, using state-of-the-art forensic and reverse engineering techniques, combined with our extensive experience as malware analysts. Through our analysis, we provide a deep understanding of the latest trends in malware, including the study of various malware families. Among other findings, we observe for the first time that: i) the level of sophistication is increasing: the new IoT malware infrastructure is borrowing sophisticated tools from Advanced Persistent Threats (APTs), ii) there is a rapid proliferation of new variants with minimal infrastructure investment, and iii) expertise and tools devised by Windows malware are permeating the IoT ecosystem.

In summary, our paper introduces a state-of-the-art analysis of Linux-based malware in the context of IoT, specifically focusing on malware prevalent in the years 2021-2023. In particular, our contributions are as follows.

1) We deploy an automated system dedicated to examining malware in IoT systems and specially engineered to examine the most recent malware samples seen in the wild at the time of writing. Our system employs a combination of well-established static and dynamic analysis, designed to establish connections between unknown samples and recognized threats.
2) Using our system, we validate recent and unidentified IoT malware samples, characterizing their behaviors. Subsequently, we present an extensive analysis of the most recent trends based on an in-depth study of several malware families.
3) We elucidate the impact of architecture on malware propagation, providing valuable insights into the dynamics influencing the recent spread of malicious activities. Our study deliberately separates Linux-based IoT malware exploration from the predominant IoT architectures associated with malware in the wild.

The full version of this paper (with a full description of the experiments) was published in [1].

Acknowledgments. This research was supported in part by grant PID2023-151467OA-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, by grants TED2021-132900A-I00 and TED2021-131115A-I00, funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, by grant *Proyecto Estratégico de Ciberseguridad – CIBERSEGURIDAD EINA UNIZAR*, funded by the Spanish National Cybersecurity Institute (INCIBE) and the European Union NextGenerationEU/PRTR, and by grant *Programa de Proyectos Estratégicos de Grupos de Investigación*, funded by the Aragonese Government (DisCo research group, ref. T21-23R). G. Suarez-Tangil was appointed as 2019 Ramon y Cajal fellow (RYC-2020-029401-I) funded by MCIN/AEI/10.13039/501100011033 and ESF Investing in your future.

(Part of) This work was supported by the European Commission under the Horizon Europe Programme, as part of the project LAZARUS (https://lazarus-he.eu/) (Grant Agreement no. 101070303). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

References

[1] J. Carrillo-Mondéjar, G. Suarez-Tangil, A. Costin, and R. J. Rodríguez, "Exploring shifting patterns in recent iot malware," in *Proceedings of the European Conference on Cyber Warfare and Security*, vol. 23, no. 1. Academic Conferences International Ltd, 2024.

[2] D. Evans, "How the next evolution of the internet is changing everything," *Cisco Whitepapers*, 2011.

[3] T. M. Attia, "Challenges and opportunities in the future applications of iot technology," 2019.

[4] A. Girish, T. Hu, V. Prakash, D. J. Dubois, S. Matic, D. Y. Huang, S. Egelman, J. Reardon, J. Tapiador, D. Choffnes *et al.*, "In the room where it happens: Characterizing local communication and threats in smart homes," in *Proceedings of the 2023 ACM on Internet Measurement Conference*, 2023, pp. 437–456.

[5] M. Ibrahim, A. Continella, and A. Bianchi, "Aot - attack on things: A security analysis of iot firmware updates," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 2023, pp. 1047–1064.

[6] A. Costin, J. Zaddach, A. Francillon, and D. Balzarotti, "A {Large-scale} analysis of the security of embedded firmwares," in *23rd USENIX security symposium*, 2014.

[7] A. Costin, A. Zarras, and A. Francillon, "Automated dynamic firmware analysis at scale: a case study on embedded web interfaces," in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, 2016.

[8] E. Cozzi, M. Graziano, Y. Fratantonio, and D. Balzarotti, "Understanding linux malware," in *IEEE Symposium on Security and Privacy (SP)*, 2018.

[9] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the mirai botnet," in *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, Aug. 2017, pp. 1093–1110. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis

[10] J. Carrillo-Mondéjar, H. Turtiainen, A. Costin, J. L. Martínez, and G. Suarez-Tangil, "Hale-iot: Hardening legacy internet of things devices by retrofitting defensive firmware modifications and implants," *IEEE Internet of Things Journal*, vol. 10, no. 10, pp. 8371–8394, 2023.

[11] B. Zhao, S. Ji, W.-H. Lee, C. Lin, H. Weng, J. Wu, P. Zhou, L. Fang, and R. Beyah, "A large-scale empirical study on the vulnerability of deployed iot devices," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 3, pp. 1826–1840, 2022.

[12] R. Li, Q. Li, Y. Huang, W. Zhang, P. Zhu, and Y. Jiang, "Iotensemble: Detection of botnet attacks on internet of things," in *Computer Security – ESORICS 2022: 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26–30, 2022, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 569–588.

[13] A. A. Al Alsadi, K. Sameshima, J. Bleier, K. Yoshioka, M. Lindorfer, M. van Eeten, and C. H. Gañán, "No Spring Chicken: Quantifying the Lifespan of Exploits in IoT Malware Using Static and Dynamic Analysis," in *Proceedings of the ACM on Asia Conference on Computer and Communications Security (ASIACCS)*. Association for Computing Machinery, 2022.

[14] A. Costin and J. Zaddach, "IoT malware: Comprehensive survey, analysis framework and case studies," *BlackHat USA*, 2018.

[15] J. Carrillo-Mondéjar, J. Martínez, and G. Suarez-Tangil, "Characterizing linux-based malware: Findings and recent trends," *Future Generation Computer Systems*, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X19325002

[16] E. Cozzi, P.-A. Vervier, M. Dell'Amico, Y. Shen, L. Bilge, and D. Balzarotti, "The Tangled Genealogy of IoT Malware," in *Annual Computer Security Applications Conference (ACSAC)*. Association for Computing Machinery, 2020.

[17] O. Alrawi *et al.*, "The circle of life: A Large-Scale study of the IoT malware lifecycle," in *30th USENIX Security Symposium*. USENIX Association, 2021. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/alrawi-circle

TRAKR: A framework to collect, process and analyze diverse APT files and reports

Alejandro Barreiro-Morante, Daniel García-Algora, Alejandro Carlos del-Rio-Álvarez, Jorge Blasco
Department of Computer Systems, Universidad Politécnica de Madrid
Calle Alan Turing s/n, Madrid, Spain
alejandro.bmorante@upm.es, d.galgora@upm.es, ac.delrio@upm.es, jorge.blasco.alis@upm.es

***Abstract*—Advanced Persistent Threats (APTs) are a significant challenge for the cybersecurity industry, requiring continuous monitoring and analysis. Cybersecurity professionals and academics share data about these threats via datasets, Cyber Threat Intelligence (CTI) reports and automatic exchanges via APIs and other less public means. In this paper we present Threat Reconnaissance and APT Knowledge Repository (TRAKR) , a framework designed to collect, process, and analyze CTI reports alongside binary samples related to APT incidents. TRAKR integrates multiple data sources, enabling correlation between threat reports and malware artifacts. We show its initial analysis capabilities, highlighting its potential for uncovering trends within APT operations and issues in how APT data is being collected. Our initial results show TRAKR's ability to enhance APT intelligence by bridging the gap between structured reports and technical malware analysis.**

***Index Terms*—APTs , Malware, CTI.**

Tipo de contribución: *Investigación en desarrollo*

I. Introduction

Advanced Persistent Threats (APTs) are now a common element of geopolitical warfare and differ significantly from traditional malware threats. They are carried out by adversaries with plenty of resources that pursue their objectives over extended periods of time, reacting to defender's mitigation techniques and trying to keep access to compromised systems for as long as possible. Although the term APT was coined 20 years ago to allow the US military to discuss the details of the intrusion with civilian experts [1], the term APT is now widely used across academia and industry.

The rapid increase in the amount and sophistication of these attacks has led to several initiatives to share and exchange information about these attacks. These range from standards to allow automated exchange of incident data such as STIX [2] to reports being published by organisations and cybersecurity vendors. These reports provide details on information such as the impact they had on their victim's systems, the threat actor's profile, and behavioral indicators such as tactics, techniques and procedures that may be identifiable. This information proves valuable not only for preemptive purposes, but also for attribution when dealing with unknown attacks [3].

In this paper, we present Threat Reconnaissance and APT Knowledge Repository (TRAKR), a system that enables the collection, processing, and analysis of Cyber Threat Intelligence (CTI) reports and binary file samples related to APTs incidents. TRAKR collects information from a variety of sources and allows to create processing pipelines that combine information from threat reports and features extracted from malware files. TRAKR can be updated at regular intervals and can be expanded easily with new sources of information and information extractors.

II. Background

APTs have undergone substantial evolution since the term was introduced by the United States Air Force in 2006. Initially conceived to facilitate discussions regarding intrusion activities between military and civilian entities, the concept has since expanded to encompass highly sophisticated, state-sponsored cyber operations with significant geopolitical ramifications [1]. The National Institute of Standards and Technology (NIST) characterizes an APT as "an adversary with sophisticated levels of expertise and significant resources, allowing it through the use of multiple different attack vectors to generate opportunities to achieve its objectives" [4]. Typically, these objectives involve establishing prolonged access to targeted information technology infrastructures to persistently exfiltrate data or disrupt critical systems.

APTs differ fundamentally from conventional cyber threats in several dimensions. While traditional attacks often involve single actors targeting unspecified systems for immediate financial gain, APTs are executed by well-resourced, organized groups against specific organizations or governmental institutions for strategic advantages [5]. The Stuxnet attack in 2010, which targeted Iran's nuclear program by manipulating programmable logic controllers (PLCs), exemplifies this evolution—demonstrating unprecedented sophistication in a cyber weapon designed for targeted physical impact through prolonged, stealthy operation [6].

Integrating malware samples with corresponding threat intelligence can enhance APT analysis. This integration can provide context that binary analysis alone cannot offer revealing strategic intentions, attribution details, and operational patterns that characterize specific threat actors. In fact, as observed in [7], this correlation enables analysts to identify similarities between campaigns and develop effective protective measures against similar attacks.

The analysis of APT attack vectors must extend beyond traditional executable files. Document-based malware, mobile applications, and fileless components have become increasingly prevalent in sophisticated attacks [8]. These vectors often serve as initial infection mechanisms, leveraging widespread use of office documents and inherent vulnerabilities in common applications. Processing these diverse file types provides additional correlation points between seemingly disparate campaigns and enhances detection capabilities against evolving APT methodologies.

III. TRAKR

This section presents TRAKR, a system designed for the integration and examination of APT malware samples along with their corresponding threat intelligence narratives. Figure 1 provides an overview of TRAKR's pipeline. It employs a multi-faceted approach to data collection, processing, and analysis, establishing correlations between different cybersecurity information sources. Contrary to most academic datasets, TRAKR is frequently updated with new reports and binary samples.

A. Crawling

The crawling component collects cyber threat intelligence and binaries from heterogeneous sources. TRAKR employs specialized crawlers adapted to each intelligence source's unique structure and accessibility constraints. We classify the data sources of our crawling component in four types: datasets, APIs, web sources, and proprietary sources. Table I provides a summary of the data sources available in the current version of TRAKR. The following section elaborates on the specifics of each of these.

1) Datasets: TRAKR combines academic and dynamic online datasets to ensure analytical rigor and relevance to current threats. Academic sources—characterized by their structural consistency—provide 33,976 samples [9, 10, 11, 12], but they typically exhibit a latency that limits their utility for tracking emerging threats (i.e., they tend to be static snapshots taken at a specific point in time). To address this, TRAKR integrates live sources such as VX-Underground[1], contributing 35,657 binaries and 2,102 reports[2] that reflect emerging APT techniques with minimal delay.

2) APIs: Public APIs deliver regulary updated threat intelligence through standardized query protocols. TRAKR utilizes the VirusTotal API to obtain 768 binary samples not present in other datasets, along with their corresponding metadata.

3) Web Sources: Many threat reports are published via CTI provider websites (e.g., FireEye, Mandiant, CrowdStrike) in formats designed for human consumption rather than automated parsing. Despite extraction challenges, these sources remain critical as they contain nuanced intelligence often overlooked in structured sources. TRAKR collects reports from 14 web sources, totaling 22,784 reports.

4) Proprietary Sources: TRAKR can incorporate proprietary datasets subject to confidentiality constraints, enriching the framework with specialized knowledge typically unavailable in public domains.

Through the integration of these diverse data sources, TRAKR establishes a basis for correlating disparate pieces of threat intelligence. In total, TRAKR includes 23 different sources of information, 6 of which are static and 17 dynamic. Table I provides an overview of these data sources from which reports and binaries are obtained.

B. Data Processing

The processing pipeline of TRAKR consists of two parallel processes (one for threat reports and another for binary files) that converge in a unified correlation engine.

[1]https://vx-underground.org/apt

[2]As of March 10, 2025.

Table I
SUMMARY OF DATA SOURCES USED BY TRAKR AS OF MARCH 2025.

Name	Type	# binaries	# reports
APTMalware [9]	Dataset	3594	-
ADAPT it! [10]	Dataset	6134	-
APT-class [11]	Dataset	20242	-
vx-underground	Dataset	35657	2102
dAPTs et [12]	Dataset	3238	-
VirusTotal	API	768	-
APT Cybercampaign	Dataset	-	1585
APT_REPORT	Dataset	-	340
APTs -database	Dataset	-	419
APTnotes	Dataset	-	688
ORKL Archive	CTI Website	-	13267
Microsoft	CTI Website	-	106
Kaspersky Securelist	CTI Website	-	310
Blackberry	CTI Website	-	286
Cyber Operations Tracker	CTI Website	-	1608
Sentinel Labs	CTI Website	-	34
Talos intelligence	CTI Website	-	57
The DFIR Report	CTI Website	-	164
Checkpoint research	CTI Website	-	397
Cyberseason	CTI Website	-	100
Cyberhint	CTI Website	-	36
Labs k7	CTI Website	-	20
ETDA Encyclopedia	CTI Website	-	4481

1) Reports: Reports are obtained in different forms depending on their source (pdf or html files). We extract the textual information via Python libraries. Some raw texts are obtained from their web source parsing the request's result with *BeautifulSoup*, while others required downloading .pdf files. These files were parsed using the *pdfplumber* library in order to extract the raw text they contain.

Indicators of Compromises (IoCs) are extracted using *iocsearcher* [13], which fetches relevant information by seeking patterns with regular expressions, like IP addresses, hashes, domain names, etc. Both the date the report was collected and its URL are also stored as attributes for traceability and reproducibility purposes.

2) Binary Files: The diverse data sources ingested by TRAKR can yield duplicate binaries from multiple origins. To address this, TRAKR employs a unified schema that uses cryptographic hashes as unique identifiers and normalizes heterogeneous metadata—including contextual attributes, attribution details, and malware labels—into a single record that preserves all available intelligence while eliminating redundancy.

Our framework leverages the VirusTotal API for multi-engine detection reports and employs AVClass2 [14] for taxonomic categorization, which normalizes vendor labels to reduce naming convention inconsistencies. For robust file type identification, we use Magika [15], which has demonstrated high accuracy in recent research [8] analyzing various file formats in APT operations.

TRAKR extracts indicators across the full spectrum of malware artifacts in APT operations, including Windows executables, document-based threats, mobile applications, and fileless malware components. This comprehensive approach reflects the evolving tactics of contemporary APT campaigns, where non-executable files like Microsoft Office documents often serve as initial infection vectors due to their widespread use and inherent vulnerabilities [8].

a) Executable Binary Analysis: TRAKR employs *capa* [16] to identify malware capabilities by analyzing code struc-

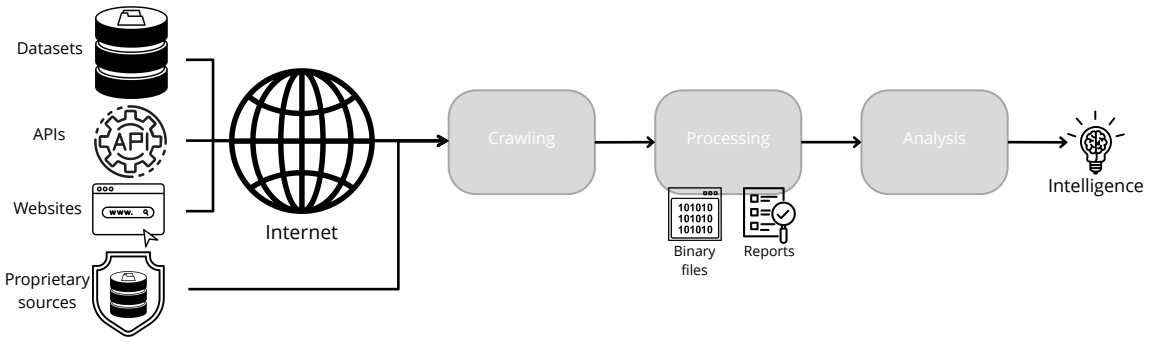


Figure 1. Overview of TRAKR

tures and API references without execution. *Radare2* [3] extracts additional indicators such as entry points and imported libraries, while *Floss* [4] deobfuscates strings, which are further processed with *iocsearcher* [13] to extract and categorize IoCs.

b) Document-based Analysis: Documents are a prevalent form of infection, TRAKR incorporates *Oletools* [17] to extract information from Microsoft *Oli2*, *RTF*, and *OpenXML* file formats, detecting features like encryption and embedded VBA macros.

c) Compressed Files: On numerous occasions, a malicious file will include other files embedded within. TRAKR extracts the files contained within others for a subsequent analysis. In its current version, TRAKR has specialized extractors for *zip* files and derivatives, and for document files from the Office suite such as *docx*, *pptx* or *odt* files, among others.

C. Analysis

TRAKR's processing component transforms raw binary files and CTI reports into structured, correlatable data. Its analysis component integrates these sources to link binaries with CTI reports, individually and together. This section introduces TRAKR's initial analsis capabilities.

a) Reports: TRAKR aggregates and analyzes CTI reports from a diverse set of sources, regarding threat actor activity, tactics, and observed incidents. By normalizing and cross-referencing information from multiple reports, TRAKR can help identify recurring patterns, correlate IoCs between campaigns, and detect changes in adversary behavior over time. The centralization of reports from both structured and unstructured sources can be advantageous in obtaining a detailed assessment of the threat landscape, enhancing the ability to track long-term trends in APT operations.

b) Malware Files: TRAKR is being developed to support the creation of binary similarity data structures that can cross-relate different malware samples based on shared components. As a work in progress, we are exploring the application of deep learning techniques to identify potential code fragments, libraries, or development patterns shared within and between APT campaigns. This ongoing effort aims to facilitate the discovery of previously unrecognized connections between seemingly disparate threat actors. In parallel, we are investigating the use of compilation timestamps to construct chronological patterns of development activity. However, this must be approached with caution, as such timestamps can be manipulated by sophisticated attackers to hinder accurate attribution.

c) Cross-correlation: TRAKR performs cross-referencing between cryptographic hashes mentioned in threat reports and those present in malware repositories, establishing concrete links between samples and their contextual descriptions. Beyond basic hash matching, the framework allows to derive intelligence from the collected data through the different processing steps detailed in the previous component.

TRAKR utilizes *capa* to align capabilities extracted from binary files with the MITRE ATT&CK framework, thereby standardizing behavioral annotations within a unified taxonomy. This correlation verifies the observed binary activities against the TTPs described in threat reports, allowing new insights to be derived from the elaboration of those reports and the malware they refer to.

In addition, TRAKR integrates IoCs from both threat reports and malware samples to create sets of indicators for identified campaigns or actors. By examining relationships between different IoCs types, the system can identify infrastructure patterns that may not be evident when examining indicators in isolation.

IV. Early Results

This section provides some early results obtained by applying the analysis components of TRAKR on data obtained up to March 10th, 2025.

A. Binary Dataset Overview

Our analysis of malware samples collected through TRAKR reveals interesting notes about the current state of APT malware collection. After applying the de-duplication process described in Section III, we mapped the distribution of samples across our source repositories to identify patterns of overlap and uniqueness.

Figure 2 presents a visual representation of the intersection between malware samples collected from various repositories. The diagram illustrates an interesting phenomenon: Despite analyzing approximately 47,821 total samples, the majority appear uniquely in individual repositories, with a small overlap between sources.

This limited intersection, 27 samples across all five repositories, suggests that current APT malware collection efforts

[3] https://github.com/radareorg/radare2
[4] https://github.com/mandiant/flare-floss

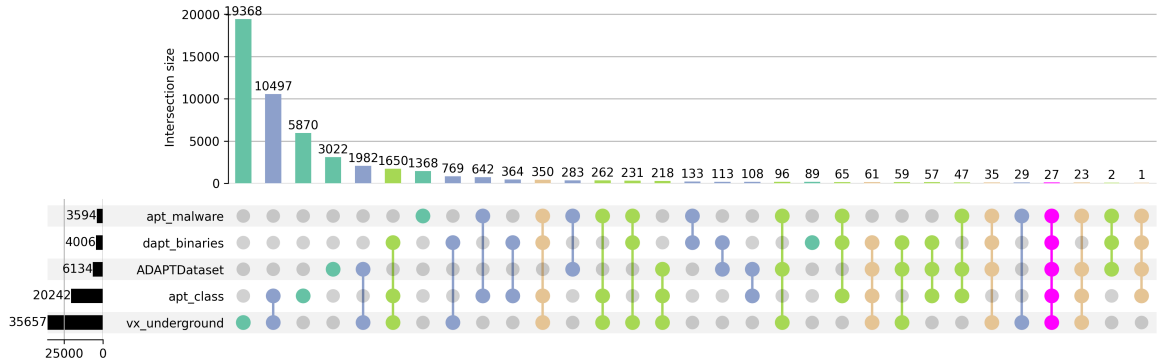


Figure 2. Venn diagram illustrating the intersection of malware files collected from various sources by TRAKR.

operate in relative isolation, with each repository capturing distinct segments of the threat landscape. The substantial number of unique samples per source indicates divergent collection methodologies and potentially different targeting priorities among cybersecurity researchers and organizations.

The minimal overlap demonstrates that relying on any single source, regardless of its reputation or size, would result in substantial blind spots in APT coverage. This supports TRAKR's multi-source integration approach, synthesizes disparate data repositories.

B. Malware Taxonomic Analysis

1) File Type Distribution: Analysis of file types in our dataset reveals diverse attack vectors employed in APT campaigns. As shown in Table II, Windows executable files (*pebin*) constitute 73.08% of collected samples, reflecting Windows' prevalence in enterprise environments.

Table II
DISTRIBUTION OF THE 8 MOST COMMON FILE TYPES IN MALWARE SAMPLES COLLECTED BY TRAKR

File Type (Magika)	# Files	Proportion
pebin	34948	73.08%
apk	2253	4.71%
doc	1007	2.11%
docx	926	1.94%
rtf	920	1.92%
elf	848	1.77%
xls	722	1.51%
zip	714	1.49%

Document formats (*rtf*, *doc*, *docx*, and *xls*) collectively represent a significant portion of the dataset, confirming their increasing use as sophisticated attack vectors in APT operations, consistent with recent findings [8]. The presence of mobile malware (*apk*), Linux executables (*elf*), and compressed archives (*zip*) demonstrates the multi-platform nature and complex deployment mechanisms of modern APT campaigns.

This diversity underscores a critical requirement for APT analysis frameworks: analytical capabilities must extend beyond conventional executable analysis to include specialized processing for document-based, mobile, and other non-traditional vectors to avoid critical blind spots in detection and characterization efforts.

2) APT Group Attribution Distribution: Analysis of APT group attributions across our dataset reveals different patterns in the threat landscape captured by TRAKR. As shown in Table III, APT17, a threat actor attributed to China, represents the largest proportion of samples at 7.79%. This is followed by SIG45 (5.59%) and the North Korean attributed actor Lazarus Group (3.75%).

Table III
DISTRIBUTION OF TOP 10 APT GROUPS ASSOCIATED WITH MALWARE FILES COLLECTED BY TRAKR

APT Group	# Files	Proportion	Country
apt17	2137	7.79%	China
sig45	1533	5.59%	Unknown
lazarus group	1028	3.75%	North Korea
sig9	749	2.73%	Unknown
sig17	748	2.73%	Israel
Gorgon Group	741	2.70%	Pakistan
sig25	683	2.49%	South Korea
oceanlotus	668	2.44%	Vietnam
Lazarus	516	1.88%	North Korea
Gamaredon	446	1.63%	Russia

The attribution distribution highlights two challenges in APT analysis. First, inconsistent naming conventions appear throughout the dataset, exemplified by "Lazarus Group" and "Lazarus" being treated as distinct entities despite referring to the same threat actor. Second, a substantial portion of the samples are associated with groups of uncertain origin (e.g. SIG45, SIG9), reflecting the difficulties in attribution.

Future work will improve APT group and country attribution by integrating additional data sources to develop an ontology system that enables more accurate and consistent identification of threat actors and their respective countries.

C. Cross-Dataset Correlation Results

From our CTI repository containing 26,000 scraped reports, we extracted 111,649 unique file hashes (40,310 MD5, 19,913 SHA1, and 51,426 SHA256 hashes). When correlated against our dataset, containing 51,257 file samples, we identified intersections of 12,208 matching MD5 hashes (30.3% of our repository's MD5 collection), 4,829 matching SHA1 hashes (24.3% of SHA1s), and 12,203 matching SHA256 hashes (23.7% of SHA256s). It is reasonable to expect reports not mentioning a large part of the hashes we obtain from files. This is because of two main reasons. First, reports tend to

mention the most relevant files observed within an attack. Second, we perform unpacking of several filetypes, resulting in additional files that may not have been included within reports. A less expected result, is the large number of hashes of each kind that we were not able to map to files within our binary dataset. This could be because these hashes refer to benign tools being used during APT attacks (also known as living-off-the-land [18]. We plan to investigate this behavior in more detail in the future.

V. Conclusions

In this paper we present TRAKR, a framework for collecting, processing and analyzing APT-related cyber threat intelligence reports and binary files. TRAKR's adaptable architecture supports data from 23 sources and analysis of 13 different file types beyond executables, with regular updates to maintain currency. Our initial analysis reveals significant fragmentation in APT research, with datasets largely operating in isolation. Approximately one-quarter to one-third of malware specimens discussed in threat reports are represented in common malware repositories, while also highlighting the significant volume of potentially novel or underreported threats documented exclusively in our intelligence corpus. This highlights the need for comprehensive frameworks that integrate diverse sources. Future work will focus on expanding TRAKR's data collection capabilities and developing new analytical methods to derive insights from the collected intelligence.

Acknowledgments

This work has been supported by the MATM project (C127/23), with the collaboration of the Spanish Institute for Cybesecurity (INCIBE), and the Recovery, Transformation and Resilience Plan funded by the European Union (Next Generation).

References

[1] I. Jeun, Y. Lee, and D. Won, "A practical study on advanced persistent threats," in *Computer Applications for Security, Control and System Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 144–152.

[2] O. C. T. I. C. T. Committee, "Stix version 2.1: Structured threat information expression," 2021, accessed: 2025-03-17. [Online]. Available: https://docs.oasis-open.org/cti/stix/v2.1/stix-v2.1.html

[3] Z. Li, J. Zeng, Y. Chen, and Z. Liang, "Attackg: Constructing technique knowledge graph from cyber threat intelligence reports," in *European Symposium on Research in Computer Security*. Springer, 2022, pp. 589–609.

[4] National Institute of Standards and Technology. Apt definition(s). Accessed: 2025-03-19. [Online]. Available: https://csrc.nist.gov/glossary/term/apt

[5] P. Chen, L. Desmet, and C. Huygens, "A study on advanced persistent threats," in *Communications and Multimedia Security: 15th IFIP TC 6/TC 11 International Conference, CMS 2014, Aveiro, Portugal, September 25-26, 2014. Proceedings 15*. Springer, 2014, pp. 63–72. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-662-44885-4_5

[6] M. Baezner and P. Robin, "Stuxnet," ETH Zurich, Tech. Rep., 2017. [Online]. Available: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/200661/1/Cyber-Reports-2017-04.pdf

[7] B. Tang, J. Wang, Z. Yu, B. Chen, W. Ge, J. Yu, and T. Lu, "Advanced persistent threat intelligent profiling technique: A survey," *Computers and Electrical Engineering*, vol. 103, p. 108261, 2022.

[8] A. Saha, J. Blasco, and M. Lindorfer, "Exploring the malicious document threat landscape: Towards a systematic approach to detection and analysis," in *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2024, pp. 533–544.

[9] C. Boot, A. C. Serban, and E. Poll, "Supervised learning for state-sponsored malware attribution."

[10] A. Saha, J. Blasco, L. Cavallaro, and M. Lindorfer, "Adapt it! automating apt campaign and group attribution by leveraging and linking heterogeneous files," in *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, 2024, pp. 114–129.

[11] J. Gray, D. Sgandurra, L. Cavallaro, and J. Blasco Alis, "Identifying authorship in malicious binaries: Features, challenges & datasets," *ACM Computing Surveys*, vol. 56, no. 8, pp. 1–36, 2024.

[12] G. Laurenza and R. Lazzeretti, "daptaset: A comprehensive mapping of apt-related data," in *Computer Security: ESORICS 2019 International Workshops, IOSec, MSTEC, and FINSEC, Luxembourg City, Luxembourg, September 26–27, 2019, Revised Selected Papers 2*. Springer, 2020, pp. 217–225.

[13] J. Caballero, G. Gomez, S. Matic, G. Sánchez, S. Sebastián, and A. Villacañas, "The rise of goodfatr: A novel accuracy comparison methodology for indicator extraction tools," *Future Generation Computer Systems*, vol. 144, pp. 74–89, 2023.

[14] S. Sebastián and J. Caballero, "Avclass2: Massive malware tag extraction from av labels," in *Proceedings of the 36th Annual Computer Security Applications Conference*, 2020, pp. 42–53.

[15] Y. Fratantonio, E. Bursztein, L. Invernizzi, M. Zhang, G. Metitieri, T. Kurt, F. Galilee, A. Petit-Bianco, and A. Albertini, "Magika content-type scanner," *Magika Content-type Scanner*, 2023.

[16] W. Ballenthin, M. Raabe, F. Team *et al.*, "capa: Automatically identify malware capabilities," *Fire Eye Threat Research Blog*, 2020.

[17] P. Lagadec, "oletools: Python tools to analyze microsoft ole2 files (structured storage, compound file binary format) and ms office documents," 2025, accessed: 2025-03-03. [Online]. Available: https://github.com/decalage2/oletools

[18] F. Barr-Smith, X. Ugarte-Pedrero, M. Graziano, R. Spolaor, and I. Martinovic, "Survivalism: Systematic analysis of windows malware living-off-the-land," in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 1557–1574.

A review of: Understanding Crypter-as-a-Service in a popular underground marketplace

Alejandro de la Cruz
Universidad Carlos III de Madrid
Madrid, Spain
alcruza@pa.uc3m.es

Sergio Pastrana
Universidad Carlos III de Madrid
Madrid, Spain
spastran@inf.uc3m.es

***Abstract*—*Crypters* transform a binary M into M' to avoid detection from Anti Viruses (AVs from now on) through a series of modifications, obfuscations and encryptions. The goal is to remain fully undetected, or FUD in the hacking jargon, while maintaining its (often malicious) functionality. The Crypter-as-a-Service model is a popular activity among the commoditization in cybercrime, due to the increased sophistication of detection mechanisms. In this business model, customers receive an initial crypter which is soon updated once becomes detected by anti-viruses. This paper provides the first study on an online underground market dedicated to Crypter-as-a-Service. We compare the most relevant products in sale, analyzing the existent social network on the platform and comparing the different features that they provide.**

***Index Terms*—Crypter, Cybercrime, Underground Forums**

Contribution type: *Published research [1]*

I. Introduction

Crypters are a type of packer that cryptographically modifies a binary to evade antivirus engines. They are often traded in underground forums. Since they need to be updated once antiviruses detect them, the Crypter-as-a-Service (CaaS) model emerged which provides such updates as required. Despite extensive research on cybercrime [2], there is a gap of studies focused on the CaaS ecosystem, from the marketing and operation perspective. Accordingly, in this work we address the following research questions: (1) What is the prevalence of crypters being traded and sought? (2) Is there any specialization?, or do crypters provide similar features? (3) What is the Social Network of users engaged in the trading of forums?.

We analyze a dedicated marketplace of crypters traded in HackForums. We analyze common properties, the techniques used to advertise their products and attract new customers, and the differences between the most popular ones. We also analyze the social network formed by the actors involved in creating, buying and interacting with the multiple products available. Overall, the contributions of this paper are: (1) We describe the CaaS model and provide the first quantitative analysis of this ecosystem in a dedicated forum. (2) We describe our custom crawler for data collection method, that allows us to obtain a total of 1,492 threads or posts and 128,384 comments in those posts along with the information of 17,751 users. (3) We conduct an analysis of the collected data, including the top100 crypters being sold, and the social network analysis of the marketplace to find market niches, the most relevant users and differences between products being sold.

We open-source the crawler, the analysis scripts and an extended version of this work with additional analyses [1].

II. Methodology

We use a field study leveraging data collected from the marketplace for crypters in HackForums. We crawl posts, which might include a pamphlet, and descriptions of the main characteristics, price, etc. of the crypter, along with comments from other users. Our crawler is designed to solve common challenges in the collection of underground forum data [3]. From each post, we extract the original post content, the image (pamphlet) used to announce the crypter (if any), and all the replies (posts) of the thread. These interactions between users and posts allow to visualize the social network of crypter producers and consumers. Overall, we collected 1,492 posts from 279 different users and 128,384 comments in those posts from a total of 15,745 users. We conducted our crawling on April 2023, collecting historical data that spans for 13 years.

We leverage this dataset to compare the published products, extracting common and different features, and analysing the evolution of this technology. We also rely on social network analysis to understand the users and interactions involved in the subforum, which enables analyses such as the types of users in the forum and most relevant crypters.

III. The Crypter-as-a-Service (CaaS) model

The Crypter-as-a-Service paradigm is based on other crime-as-a-service models that offer a malicious service to potential clients. The commoditization of such products and services fosters the growth of cybercrime. This model involves several roles, which might be operated by the same individual, or multiple. Using our quantitative analysis, we spot multiple products from actors claiming not to be the developer of the crypter, but *reseller*. Besides these, the three most common roles are: (1) The *developer* team, in charge of programming and updating the main technical components of the crypter. (2) The *commercial* team advertises the service. Its main goal is to attract customers and also to provide customer support. (3) The *finance* team creates and maintains the payment platform. This is often based on cryptocurrencies (e.g., Bitcoin, Ethereum or Monero), or by other online payment systems popular in criminal businesses, such as PayPal or Amazon Gift Cards. It must keep the payments anonymous, and also is responsible for the money laundering.

IV. Forum Analysis

Properties. We analyse the most recurrent words related to properties of the crypter. These include *"FUD"*, ratings of

the product, users asking for free *"trials"*, the type of stub used *"private"*, AVs *"norton"* and contact info *"discord"*.

Evolution. Next, we analyze the activity on the forum and its evolution across time. Since 2011 there has been a moderate decrease in crypter activity, first by a cyberattack leading to leaked credentials from nearly 200,000 users. Additionally, as explored by Bhalerao et al. [4], in 2014 malware obfuscation alternatives became more frequent and crypters lost their popularity.

Expertise. The last approach is extracting the number of posts asking for help, to understand the prevalence of newcomers and actors seeking obfuscation for their cybercriminal goal. A simple –yet effective– way to achieve this classification is by using keywords in the context of asking for help such as "help", "need" or "advice". 24.87% of the total amount of threads are asking for something instead of selling a product.

Cutting-edge products. We manually analyse the top 100 cryters by number of views. Out of these, 11 of them were currently closed or paused at the time of the analysis, and only 15 have been active since 2020, which is in accordance with the timeline analysis provided in the previous section and the overall decline in popularity. Regarding their providers, there are 13 users that have created (or advertised) more than one. This users resemble the role of a reseller or "commercial".

Technical properties. Crypters offer similar functionalities (custom startup, virtual machine and sandbox detection, etc.), payment methods (often reling on cryptocurrencies and anonymous payment methods such as PerfectMoney). Most announce 24/7 customer support, reply to comments in the forum and update the original post of the thread quite often to appear active. As they are products with potential misuse being sold in a public website, they often disengage from any responsibility with a disclaimer, typically forbidding their usage for illegal activities.

Architecture. There are three types of stubs: private, which allegedly means a different one for every buyer; standard, which is shared among customers (or by every N users); and the unique stub generator, which is a combination of both. A private stub implies more work than a standard one, which is reflected in the price, being up to 10 times more expensive. The less users using the same stub, the lower the probability of it being fingerprinted and detected by AVs, which is a clear advantage.

Demographics. We analyze the types of users based on their forum activity: number of posts they create, prevalence of users than only leave comments, etc. A vast majority of the users (98.23%) only comment, instead of publishing threads, which is expected in a forum-like marketplace. Two-thirds of the users only interact with one post, which suggests the general tendency for low interaction with the sub-forum apart from buying and help-seeking. Lastly, we have the "creators", i.e., users that publish one or more crypters, where the most common case is users that create one post and comment on several others.

Social Network. To better reflect the interactions of users with posts, we have created a weigthed graph of the social connections of posts and users within the top 100 posts, with two type of edges, i.e., creation and commenting of posts. The weight reflects the number of comments of a given user in a given post. The graph is fully interactive and available online [5]. The graph shows that this is a homogeneous network, suggesting that the community is equally distributed, with no particular niches and specialized topics. The degree centrality of the network provides the "popularity", which we compare with the actual "reputation" of users in the forum. We, however, observe few relation between the in and out degree in the graph and the popularity within the forum, showing that the crypter marketplace is an isolated environment. Those users with most degree in the marketplace have very low forum reputation and vice versa. On the other hand, the eigenvector centrality (i.e. the relevance of a node from the importance of its neighbor nodes) offers more valuable results. The top five nodes with highest eigenvector values are present in the top posts that are still active since 2020. Highlighting that ultimately the most important factor popularity-wise is creating a popular product and mantaining it over time.

V. Conclusions

In this paper, we study the Crypter-as-a-Service ecosystem in HackForums. Most of the advertised products and services offer similar features in their adverts, with similar properties, payment methods, customer support and close prices. The differentiating part is the way each crypter creates the stub (i.e., the encryption process), and also the customer support, (i.e., the treatment given to customers once the service or product is delivered). Finally, the social network analysis shows that most users do not engage actively in the market, with few actors acting as (re)sellers or developers. These insights may assist law enforcement, since it suggest that an effective tactic would be to focus on the common techniques (e.g., for antivirus industry) and actors (e.g., for law enforcement officers).

Acknowledgments

This paper was partially supported by INCIBE grant APAMCiber within the framework of the Recovery, Transformation and Resilience Plan funds, financed by the European Union-NextGenerationEU, and by grant TED2021-132170A-I00 from the Spanish Ministry of Science and Innovation, funded by MCIN/AEI/10.13039/501100011033, and the European Union-NextGenerationEU/PRTR.

References

[1] A. De La Cruz and S. Pastrana, "Understanding crypter-as-a-service in a popular underground marketplace," in *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2024, pp. 85–90.

[2] J. Hughes, S. Pastrana, A. Hutchings, S. Afroz, S. Samtani, W. Li, and E. Santana Marin, "The art of cybercrime community research," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–26, 2024.

[3] K. Turk, S. Pastrana, and B. Collier, "A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments," in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2020, pp. 428–437.

[4] R. Bhalerao, M. Aliapoulios, I. Shumailov, S. Afroz, and D. McCoy, "Towards automatic discovery of cybercrime supply chains," 2018.

[5] A. de la Cruz. (2023, Jun.) Hackforums crypter social network. [Online]. Available: https://paquitoelchocolatero.github.io/HFCrypterAnalysis/social-network/canvas.html

A Review Of: Reversing the Virtual Maze - An Overview of the Technical and Methodological Challenges for Metaverse App Analysis

Alfonso Rodriguez Barredo-Valenzuela[†‡], Sergio Pastrana Portillo[‡], Narseo Vallina-Rodriguez[†]
Guillermo Suarez-Tangil[†]
[†]*IMDEA Networks Institute*, [‡]*Universidad Carlos III de Madrid*

***Abstract*—The Metaverse, powered by Extended Reality (XR) technologies such as Virtual Reality, Augmented Reality, and Mixed Reality, is rapidly evolving, offering new opportunities but also introducing significant security and privacy challenges. The proliferation of diverse platforms, devices, and development frameworks creates a highly fragmented ecosystem, complicating large-scale application analysis. In addition, the integration of novel sensors increases the risk of user tracking and data leaks. This work presents an experience-based vision of the key technical and methodological challenges in Metaverse app security and privacy analysis that need to be tackled to effectively address emerging threats.**

***Index Terms*—Metaverse, Threat Analysis, Security, Privacy**

Type of contribution: Research already published [1]

I. Introduction

The Metaverse is made up of three technologies: Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). Applications extend beyond gaming to sectors like health and culture. This growing field is projected to have a financial impact of $1.5 trillion by 2030, with major tech companies such as Mozilla, Google, and Netflix investing in VR applications. The Metaverse raises significant security, privacy, and safety concerns. However, there is limited literature on these issues compared to other fields like smartphones. The lack of market consolidation has led to a fragmented landscape with diverse platforms, devices, and tailored solutions, complicating risk detection and mitigation. This complexity is further exacerbated by the unique characteristics of Metaverse technologies, such as sensors, controllers, and 3D interactions, which arises new XR-specific risks, e.g., room fingerprinting, virtual asset theft, or virtual harassment.

Previous work on S&P issues in the Metaverse [2], [3], [4], [5], [6], [7], [8], discuss technical, ethical, social, or economic problems. However, these works offer an incomplete overview of the threat landscape, ignoring key challenges that need to be addressed to procure practical security risks over real-world applications. Thus, based on our own experience and a complete literature review, we conduct a qualitative study to understand existing methodological and technical challenges preventing the analysis of the Metaverse ecosystem at scale. Specifically, we answer three key research questions. **RQ1:** What are the emerging security and privacy risks that current threat models fail to adequately capture?; **RQ2:** What techniques can be adopted from other domains to automate the analysis of Metaverse applications?; and **RQ3:** What are the key challenges in developing robust and scalable techniques for the automatic analysis of Metaverse applications?

Our study goes a step further from previous work [9] by not only considering the potential S&P threats but also delving into concrete issues related to the empirical analysis of applications in the Metaverse. Thus, our goal is to ease the understanding of the entire ecosystem and then pinpoint specific issues and obstacles associated with analyzing XR software, either with dynamic or static analysis methods.

II. Background

The Metaverse is a complex ecosystem comprising a variety of platforms, devices, operating systems, and development tools. XR devices are either standalone, capable of running applications independently, or non-standalone, requiring an external device for processing. The market has shifted toward standalone devices, which now account for the majority of XR hardware. This transition has driven the proliferation of dedicated operating systems, mostly Android-based systems.

The development of Metaverse applications follows a structured cycle. Developers rely on both platform-specific SDKs, which provide native functionality, and broader development SDKs integrated into engines like Unity and Unreal. OpenXR has emerged as a key industry standard, enabling developers to create applications that function across multiple platforms without requiring extensive modifications. Then, they are distributed through various marketplaces, ranging from official stores like Meta Store and Apple's App Store to third-party platforms such as Steam and SideQuest.

The new paradigm in the Metaverse introduces new risks related to user-to-user interactions within the XR environment and the cyberphysical environment. Privacy risks stem from the collection of sensitive user data, enabling user and room fingerprinting and profiling, and potential data thefts. Security threats arise from various attack vectors, including malicious software and actors, device hicjacking, and virtual extorsion, scams or harassment. The immersive nature of XR environments exacerbates these risks, making users more vulnerable to identity-based threats and behavioral tracking. Only a few efforts have empirically analyzed privacy risks in real XR applications with limited visibility (e.g., [10]). Addressing these gaps requires a more comprehensive approach that integrates practical security assessments with theoretical threat models, ensuring that privacy and security considerations keep pace with the rapid evolution of the Metaverse. Our work provides an informed analysis of key technical and methodological challenges that the community must address to effectively assess S&P threats of real-world Metaverse applications.

III. Challenges

Data Crawls: Three primary installation paradigms define how Metaverse applications are obtained and executed. First, some applications run on external devices like PCs or gaming consoles, with XR headsets serving only as controllers or display devices. Second, certain PC-based marketplaces, such as SideQuest, allow direct installation of applications onto standalone XR devices. Third, some marketplaces operate entirely within the headset itself, enabling direct downloads and execution on the same device.

There are 2 main approaches to automatically collect Metaverse applications. One involves interacting with marketplace user interfaces, requiring extensive modeling of different store layouts and installation processes to enable effective navigation. Another approach focuses on crawling apps through APIs, which demands reverse-engineering device-market interactions and bypassing server-side restrictions to access installation services. Both methods present technical challenges, but overcoming them is essential for scalable and systematic app analysis.

Static Analysis: It is hindered by platform diversity, customized operating systems, and evolving development frameworks. Many platforms modify standard OS structures, introducing unique APIs, security mechanisms, and software components that complicate traditional analysis pipelines. Understanding these architectures is essential for assessing security and privacy risks, but the variety of SDKs and cross-platform tools makes lifecycle modeling complex. Existing mobile-focused static analysis methods must be adapted to accommodate Metaverse-specific abstractions and asynchronous execution models.

Code analysis is further challenged by the use of multiple programming languages, native libraries, and obfuscation techniques that protect intellectual property. Vulnerability identification requires tracking dependencies across different SDKs, making it difficult to create a comprehensive Software Bill of Materials (SBOM) for the Metaverse ecosystem.

The fragmented nature of these platforms and development practices means that no scalable static analysis pipeline currently exists, highlighting the need for new methodologies tailored to the Metaverse landscape.

Dynamic Analysis: This is constrained by limited access to device permissions, the complexity of simulating interactive XR environments, and challenges in instrumentation. Unlike traditional platforms, Metaverse applications run on diverse operating systems with restricted access to system resources, making it difficult to monitor runtime behaviors comprehensively. Instrumentation can occur at the network, OS, or application level, but each approach faces obstacles such as anti-cheat mechanisms, signature-based authentication, and the difficulty of modifying root certificates for traffic interception. The closed nature of Metaverse platforms further complicates the analysis of encrypted communication, hindering the detection of security risks like personal data leaks.

Automating the excecution of XR applications is another significant challenge. Manual testing is impractical due to the complexity of 3D environments, continuous movement, and dynamic user interactions. Unlike mobile or desktop applications, XR experiences involve variable user behaviors, sensor inputs, and unpredictable external factors, making automated testing and fuzzing extremely difficult. Additionally, detecting harmful user behavior within Metaverse applications, such as harassment or exploitation, requires new moderation techniques beyond those used in traditional social networks. The lack of scalable tools for analyzing Metaverse applications and user interactions highlights a pressing need for novel methodologies that account for the distinct characteristics of XR environments.

IV. Conclusions

Our study highlights the critical challenges in analyzing Metaverse applications at scale, emphasizing gaps in security research, the fragmentation of XR platforms, and the lack of scalable methodologies for app collection and instrumentation. Current approaches, particularly in static and dynamic analysis, struggle with the complexity of XR applications due to proprietary architectures, extensive system dependencies, and the inability to automate interactions effectively. The Metaverse ecosystem introduces diverse security risks that are not fully addressed by existing research, necessitating the development of new techniques to assess vulnerabilities, privacy risks, and user behaviors comprehensively. Our work is the first to systematically distill these challenges, setting the foundation for advancing Metaverse security analysis.

Future research should focus on overcoming key barriers such as architecture understanding, component analysis, and automated data collection. Addressing these foundational issues will enable scalable approaches, equipping researchers with the tools necessary to find the way through the ”Virtual Maze” to effectively protect its explorers.

Acknowledgments

This project was supported by various funding bodies as acknowledged in [1]. The dissemination in JNIC is funded by PID2022-143304OB-I00 (PARASITE) funded by MICIU/-AEI/10.13039/501100011033/ and by the ERDF, EU. G. Suarez-Tangil has been appointed as 2019 Ramon y Cajal fellows (RYC-2020-029401-I) funded by MICIU/AEI/10.-13039/501100011033 and the ESF Investing in your future.

References

[1] A. Rodríguez Barredo, S. Pastrana, N. Vallina-Rodriguez, and G. Suarez-Tangil, “Reversing the virtual maze: An overview of the technical and methodological challenges for metaverse app analysis,” *IEEE MetaCom*, pp. 1–9, 2024.

[2] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, “A survey on metaverse: Fundamentals, security, and privacy,” 2022.

[3] D. Adams *et al.*, “Ethics emerging: the story of privacy and security perceptions in virtual reality,” 2018.

[4] R. Di Pietro *et al.*, “Metaverse: security and privacy issues.” IEEE, 2021.

[5] M. Vondráček *et al.*, “Rise of the Metaverse’s Immersive Virtual Reality Malware and the Man-in-the-Room Attack & Defenses,” 2023.

[6] C. Sandeepa *et al.*, “Privacy of the Metaverse: Current Issues, AI Attacks, and Possible Solutions.” IEEE, 2023.

[7] V. Nair *et al.*, “Exploring the privacy risks of adversarial vr game design,” *Proceedings on Privacy Enhancing Technologies*, 2023.

[8] T. Zhang *et al.*, “FaceReader: Unobtrusively Mining Vital Signs and Vital Sign Embedded Sensitive Info via AR/VR Motion Sensors,” 2023.

[9] G. M. Garrido *et al.*, “SoK: Data Privacy in Virtual Reality,” 2023.

[10] R. Trimananda *et al.*, “OVRseen: Auditing network traffic and privacy policies in oculus VR,” in *31st USENIX security symposium*, 2022.

Defining the Basal Attractiveness Concept for Cybercriminals

Javier García-Ochoa
Rey Juan Carlos University
Móstoles, Spain
javier.garciaochoa@urjc.es

Alberto Fernández-Isabel
Rey Juan Carlos University
Móstoles, Spain
alberto.fernandez.isabel@urjc.es

Isaac Martín de Diego
Rey Juan Carlos University
Móstoles, Spain
isaac.martin@urjc.es

Clara Contreras
Rey Juan Carlos University
Móstoles, Spain
clara.contreras@urjc.es

Romy R. Ravines
DeNexus Inc.
Boston, United States
rr@denexus.io

Ovidio López
DeNexus Inc.
Boston, United States
ol@denexus.io

***Abstract*—Companies increasingly face more frequent and diverse cyber incidents, posing significant risks to their operations and reputation. This paper proposes a novel approach to measure *Basal Attractiveness* of companies to cybercriminals based on inherent attributes that define the company. With these attributes, alongside Machine Learning techniques, this methodology aims to provide a more accurate assessment of a company's exposure to a cyber incident. The approach uses Rule Mining to identify hidden patterns in victims of cyber incidents, which are then applied to generate features for classification tasks. A Decision Tree classifier is trained to predict the likelihood of a cyber incident based on the derived Rule Mining scores. The proposed method is tested using a dataset that contains various company-specific variables. The experiments show promising results, indicating that companies with higher *Basal Attractiveness* scores are usually more likely to experience cyber incidents.**

***Index Terms*—Attractiveness, Cybersecurity, Cyber risk, Cyber incident, Rule Mining, Exposure**

Type of contribution: *Ongoing research*

I. Introduction

In recent years, companies have faced an increasing number of cyber incidents that can severely impact their operations and reputation [1]. These incidents, known as cyber attacks, are any kind of malicious activity that attempts to collect, disrupt, deny, degrade, or destroy information system resources or the information itself [1]. Cyber attacks can come in many forms, and are becoming more sophisticated and frequent, posing a growing threat to organizations regardless of size or industry.

The integration of technology into business operations makes companies highly vulnerable to cyber incidents, leading to potential economic losses, reputational damage, and legal consequences if data protection regulations are not met [2]. High-profile attacks on multinational corporations, such as Sony and Equifax, demonstrate the financial and reputational impact, with the average cost of a data breach reaching $4.88 million in 2024[2]. Therefore, understanding a company's exposure to cyber risk is essential for implementing effective mitigation strategies, whether through risk avoidance, transfer, acceptance, or reduction [3].

This paper presents a novel methodology to measure a company's *Basal Attractiveness*, focusing on stable attributes that make it an appealing target for non-targeted cyber incidents. The proposed methodology is part of the DICYME Project (2021-009025), a collaboration between Rey Juan Carlos University and DeNexus Inc., aimed at quantifying cyber risk in industrial environments through probabilistic models and real-time data analysis. By combining public and private data sources, DICYME enables continuous and updated risk assessments, supporting the identification of threats and the implementation of defensive strategies.

The paper is structured as follows. The background is described in Section II. The proposed methodology is presented in Section III. The dataset construction and characteristics are detailed in Section IV. The experiments are depicted in Section V, while upcoming developments are discussed in Section VI. Finally, Section VII concludes.

II. Background

Attractiveness measure is a modern concept in the cybersecurity domain [4]. This measure typically considers general attributes such as the country, the category, or the financial value of companies.

Additionally, Data Mining and Machine Learning techniques can significantly enhance the accuracy and efficiency of assessing a company's cyber risk and attractiveness. Machine Learning tools can analyze several relevant parameters to assign a class of cyber risk to organizations [5]. Furthermore, advanced Machine Learning models can support decision-makers by predicting and managing cyber risks more effectively [6].

Specifically, several approaches highlight rule-based methods in cyber risk measurement and management. For example, a rule-based business process mining approach was proposed to manage uncertainty and risk in security service providers [7].

Fuzzy rule sets are also used for mining association rules, applied to security event correlations and risk assessment in information security systems [8]. Association Rule Mining

[1]https://csrc.nist.gov/glossary/term/cyber_attack
[2]https://www.ibm.com/reports/data-breach

algorithms based on multi-source logs have also been used to detect cyber intrusion behaviors on different cloud computing platforms [9], and risk management issues using process mining techniques for compliance [10].

In contrast, the methodology presented in this paper delves into specific, concrete attributes that characterize a company and are relatively stable, though not entirely immutable, such as its revenue, earnings, public trading status, profitability, and number of employees. These attributes serve as fundamental indicators of the appeal of a company to potential cybercriminals, particularly those without predetermined targets seeking a specific one. The novelty lies in integrating a rule-based approach to identify patterns in historical cyber incidents and combining it with a Decision Tree classifier for predictive modeling. This dual-method approach, which leverages both association Rule Mining and supervised learning, allows for a robust estimation of cyber risk exposure, surpassing the capabilities of traditional static risk assessments.

III. Methodology

This section presents the methodology used to evaluate the *Basal Attractiveness* concept. It consists of a measure designed to capture the inherent qualities that make a company an attractive target for cybercriminals. Unlike models that rely on dynamic or external factors (e.g., current threat landscapes), this approach focuses on company-specific attributes that remain relatively stable over time also called firmographic data. They are significant for actors seeking economic gain or reputation without a specific target, as the attributes indicate where they can extract value, whether through financial exploitation, data theft, or reputational damage.

The dataset is filled using information from confirmed victims. This information is completed with similar companies (i.e., specifically sharing variables such as country and industry category with those companies that disclosed an incident). This process leads Machine Learning models to be trained with a dataset that contains companies affected and unaffected by confirmed cyber incidents (see Fig. 1 for a graphical depiction of the process).

Once the data of companies have been collected, the methodology uses a Rule Mining approach, specifically the FP-Growth algorithm [11], a well-known and efficient method for identifying frequent item sets in large datasets by packing data into a compact FP-tree structure. This algorithm was selected for its scalability, suitability, and efficiency in handling large datasets with high-dimensional attributes, compared to alternatives such as the Apriori algorithm, which can suffer from high computational costs.

The FP-tree model is used to discover patterns or associations within the dataset (see Table I), focusing on how combinations of company attributes relate to the occurrence of incidents (i.e., the model in this step must use only confirmed incidents to be trained).

Later, a Decision Tree algorithm is used to classify companies into a range $[0, 1]$. The lowest value indicates a lack of attractiveness, while the highest value indicates outstanding attractiveness. This value is called the *Basal Attractiveness* of the company (i.e., a static estimation of the probability of a cyber incident to occur). As previously stated, this approach can rely solely on publicly disclosed incidents, rather than on attacks or attempts that are unquantifiable and inherently internal to organizations.

Table I
Dataset generated with the set of rules from the Rule Mining algorithm

Attribute	Description
Rules amount	Number of rules met by the entry.
Total support	Sum of the support of all the rules met by the entry. The support is the frequency with which a set of items appears in the dataset.
Total confidence	Sum of the confidence of all the rules met by the entry. The confidence is the probability that the item set B appears in transactions that already contain the item set A.
Total lift	Sum of the lift of all the rules met by the entry. The lift measures the relationship between the joint frequency of a set of items and the expected frequency if the items were independent.
Label	Incident or non-incident.

The training process of the Decision Tree uses the incidents previously selected to obtain the set of rules and the same percentage of non-incidents. These data are processed to calculate the sum of lift, support, and confidence of each satisfied rule and the amount of these rules for each company. Then, this result is used as input in the model.

Notice that the hyper-parameters of the Decision Tree are tuned based on a grid search over a validation set. Moreover, this model also provides interpretability. This transparency is important, especially for business decision-makers who may not be familiar with complex Machine Learning models such as Neural Networks.

IV. Dataset

The dataset used in the proposal was gathered from multiple sources, capturing the specified attributes when the cyber incident occurred.

Cyber incidents have been manually extracted from dedicated public databases: European Repository of Cyber Incidents (EuRepoC)[3], Hackmageddon[4], Jam Cyber Successful Cyber Attacks Database[5], TI Safe Incident Hub[6], KonBriefing[7], CISSM Cyber Attacks Database[8], and ICS STRIVE[9]. For each cyber incident that involved a single identifiable victim, firmographic attributes were collected using specialized tools provided by DeNexus [12].

The same attributes were gathered from similar companies unaffected by a confirmed cyber incident. These companies have been identified using the same tools, but a manual verification is performed to guarantee the process's accuracy, ensuring that both the sector and country variables match. See Fig. 1 for details on the data-gathering process.

The selected variables align with state-of-the-art approaches, which mainly consider location and industry (reflecting a company's geopolitical association and criticality of operations), financial status, and human resources indicators

[3]https://eurepoc.eu
[4]https://hackmageddon.com
[5]https://jamcyber.com/discover/cyber-attacks
[6]https://hub.tisafe.com
[7]https://konbriefing.com/en-topics/cyber-attacks
[8]https://cissm.umd.edu/cyber-events-database
[9]https://icsstrive.com

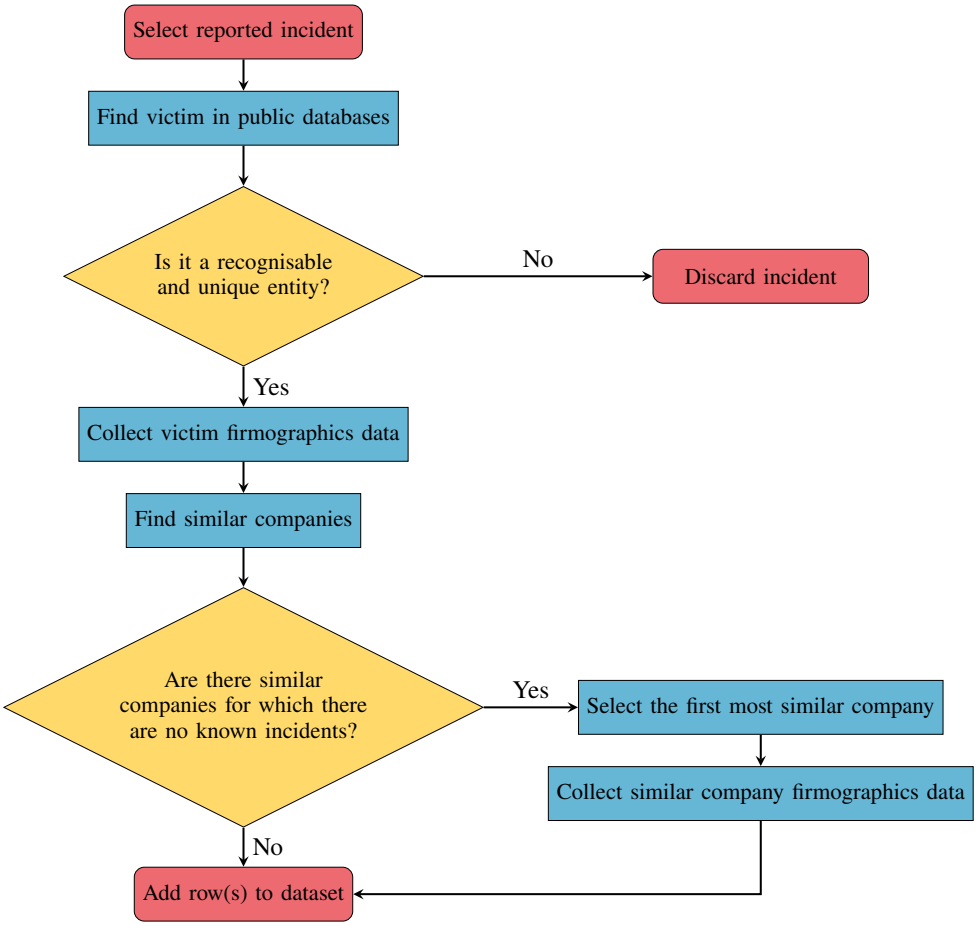


Figure 1. Working steps during the data collecting stage in the methodology.

[13], [14], [4]. Following this context, the following attributes, which remain relatively stable over time, have been chosen:

- **Country**: Companies operating in certain countries may be targeted more frequently due to geopolitical tensions, weaker cybersecurity regulations, or high economic value.
- **Sector**: Some sectors may be more attractive targets because of the sensitive nature of their data or their operational importance.
- **Revenue**: Higher revenue companies may be more attractive targets due to the potential financial gain.
- **Earnings**: Similar to revenue, higher earnings may increase a company's attractiveness to cybercriminals.
- **Publicly traded**: Public companies might be attractive due to the potential for stock price manipulation and the public availability of information.
- **Employees**: Larger companies might be more attractive because of the larger attack surface, as the human factor is one of the most common entry points or starting points for an attack.
- **Profitable**: Profitable companies might be perceived as lucrative targets.

In addition to these definitions, the dataset includes the date, entity (i.e., company name), and incident category (ransomware, data breach, denial of service, destruction, or non-incident). In this case, the categories have been limited to confirmed incidents and non-incidents. This modification simplifies the dataset and fits well with the methodology concept. It's important to note that data collection was conducted as close as possible to the moment the incident was disclosed, aiming to minimize subsequent fluctuations, such as financial losses.

Therefore, the resulting dataset has 675 observations (190 non-incidents and 485 incidents) selected between December 2023 and December 2024.

V. Experiments

This section details the evaluation of the proposed methodology to measure the *Basal Attractiveness* value of a company to cybercriminals. The experiments were achieved using the dataset previously detailed.

The dataset is divided into two parts; incidents and non-incidents. The incidents dataset is divided into training (80% i.e., 388 observations) and testing (20% i.e., 97 observations). For the sake of brevity, we will refer to training incidents as the subset of observations from the training dataset of incidents. Similarly, the non-incident dataset follows the same division process, with 80% allocated to training (i.e., 152 observations) and 20% to testing (i.e., 38 observations), ensuring consistency across both datasets for model evaluation and validation.

Based on the proposed methodology, the training incidents feed the Rule Mining algorithm, obtaining a set of rules. During the algorithm's execution, rules are filtered by setting a minimum confidence and support of 0.01, resulting in a final set of 8,214 rules. Note that numeric variables are transformed into categorical data with four discrete intervals to prevent the algorithm from being overwhelmed by a large number of integer values. Although the confidence is reasonably high, ranging from $[0.86, 1]$, the support remains very low, with only 13 rules above 0.5. However, the significance of the rules will be addressed later with the aggregation metrics and the Decision Tree.

Resulting rules suggest that firms with missing earnings data and those not publicly traded share common characteristics. A subgroup involving governments is also easily distinguishable due to its high lift, which aligns with logical antecedents such as not being publicly traded, not being aware of its earnings and revenue, and lacking profitability.

However, while the more distinct and defined patterns are visible, such as these naturally explainable relationships, many subtler, more diffuse lines stand for other relationships that are of true significance for Rule Mining. They provide underlying patterns in the data, which are the foundation for calculating the *Basal Attractiveness*.

These rules are then applied to the combined training subsets of incidents and non-incidents to calculate metrics for each row (see Table I). These transformed features are used to train a Decision Tree classifier that estimates the probability that an observation is an incident (that is, the concept *Basal Attractiveness*). Different values for maximum tree depth (5, 10, 15) and minimum samples per leaf (2, 5, 10) have been tested. The best configuration was obtained with a maximum depth of 5 and a minimum of 10 samples per leaf.

The resulting classification tree reveals how the right branch, when support is less than 101, directly leads to a leaf node characterized by a high probability of being an incident, and contains the 58% of observations. On the left branch, the model further divides entities based on the amount of satisfied rules and support. Among those divisions yields the only instance in which lift plays a role. This path contains multiple leaves, whose composition varies in terms of incident likelihood, depending on the number of rules that are satisfied by each entity. Remarkably, confidence has a wide range between 0.01 and 1.00 and relatively uniform distribution, does not improve classification performance and therefore it has not been used by the tree.

After training the Decision Tree with the calculated rules for the combination of incidents and non-incidents training subsets, the testing phase involves applying the same transformations to the remaining testing subsets of incidents and non-incidents, which are subsequently used to evaluate the performance of the decision tree (see Table II).

While the Decision Tree demonstrates strong recall in both training and testing, the overall accuracy and precision are comparatively lower. This suggests that although the model is highly effective at identifying actual cyber incidents, it struggles with differentiating between incidents and non-incidents with the same level of precision. The high recall indicates that the model is capturing most of the true incidents,

Table II
Performance Metrics of the Decision Tree Classifier

Metric	Train Set	Test Set
Accuracy	75.19%	67.41%
Precision	75.71%	71.20%
Recall	96.39%	91.75%
F1-Score	84.81%	80.18%

which is crucial in cybersecurity applications where failing to detect a real threat can have significant consequences. However, the drop in accuracy implies that the model is also misclassifying a notable proportion of non-incident cases as incidents, leading to a higher number of false positives.

Fig. 2 provides additional insight into this pattern by illustrating the distribution of cyber incidents across different levels of *Basal Attractiveness*. The bins were empirically determined to optimize separation, by analyzing the data distribution and identifying three distinct clusters: $[0, 0.4]$, $(0.4, 0.7]$, and $(0.7, 1]$. The stacked bar chart indicates that as the score increases, the proportion of cyber incidents also increases significantly. A high proportion of incidents with a high *Basal Attractiveness* suggests that this feature plays a crucial role in determining risk levels. However, the presence of non-incidents in the high category highlights the challenge faced by the model, which predicts some false positives.

In conclusion, the high recall values demonstrate the quality of the model to detect cyber incidents, while some false positives are raised. Therefore, companies with higher *Basal Attractiveness* values can be flagged for more stringent security measures, while those with lower scores might carry out additional risk assessments or compensatory controls. These insights not only optimize cybersecurity budgets by aligning investments with the organization's most vulnerable areas but also provide tangible, data-driven justification for resource allocation. This enables Chief Information Security Officers (CISOs) to present empirical evidence to boards of directors, supporting their requests for increased funding and resources to address specific threats, thus facilitating informed and strategic decision-making at the executive level.

VI. Upcoming developments

Future research will focus on addressing key limitations and enhancing the robustness of the current approach. One critical area of improvement is minimizing the presence of unknown or missing data, particularly ensuring that NAs in earnings variables are meaningful rather than artifacts of data collection inconsistencies. This is essential to prevent unintended biases that could negatively impact rule-based decision-making. Additionally, we will explore techniques to mitigate the effects of missing or ambiguous information by integrating imputation strategies or leveraging alternative data sources to infer missing values reliably.

Another significant focus is improving the representativeness of the dataset. Currently, the dataset includes many more incidents than non-incidents, which does not accurately reflect the real-world distribution, where incidents occur less frequently relative to the overall population of organizations. This fact can lead to deviations in the Decision Tree classifier since it observes a population that does not conform to reality. To address this, we will gather new data as well as implement

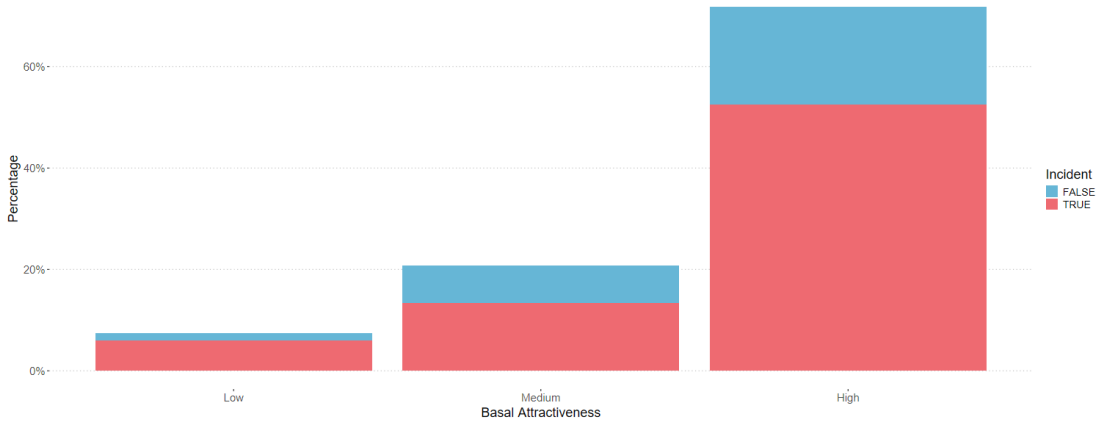


Figure 2. Distribution of the *Basal Attractiveness* scores and cyber incidents in test.

re-sampling techniques, synthetic data generation, or cost-sensitive learning approaches to better approximate real-world conditions to improve model performance.

Furthermore, we are expanding our feature set to include more dynamic variables, such as real-time threat intelligence, an entity's visibility in social media and professional networks, mentions in underground forums and the dark web, and publicly visible security posture. Incorporating these evolving risk signals will enhance the generalization and robustness of the model, enabling a more comprehensive assessment of cyber risk exposure.

As part of the evaluation process, future work will also include a comprehensive comparison with other Machine Learning models, such as ensemble methods, Random Forest, and gradient boosting algorithms. This analysis will assess the suitability of the current model by determining whether alternative approaches can establish similar or stronger relationships between the predictor variables and cyber incidents. By benchmarking performance across multiple models, we aim to validate the robustness of the proposed methodology and identify potential improvements that can be applicable in cybersecurity risk assessment and quantification.

VII. Conclusions

This work presents a novel methodology to measure the *Basal Attractiveness* value of a company by leveraging stable attributes (e.g., revenue, number of employees). The approach combines Rule Mining techniques for feature extraction with a Decision Tree algorithm to produce the final classification.

Experiments demonstrate that this combination effectively estimates the *Basal Attractiveness* value. This method aids companies in proactively managing cybersecurity resources by identifying high-risk profiles, allowing for better prioritization and mitigation of potential threats. Furthermore, integrating this methodology into cybersecurity frameworks can enhance long-term risk management strategies.

Future improvements should address data and modeling limitations to enhance accuracy, including refining data collection and expanding features to incorporate dynamic indicators like real-time threat intelligence. Additionally, comparing the Decision Tree with advanced techniques (e.g., ensemble methods) could ensure optimal performance and model robustness.

Acknowledgements

This work has been funded by the Spanish MICIU under the CPP program in the DICYME project (Ref: CPP2021-009025), partially funded by the XMIDAS project (PID2021-122640OB-I00), and supported by DeNexus Inc.

References

[1] O. K. Tosun, "Cyber-attacks and stock market activity," *International Review of Financial Analysis*, vol. 76, p. 101795, 2021.

[2] M. Chung, "Signs your cyber security is doomed to fail," *Computer Fraud & Security*, vol. 2020, no. 3, pp. 10–13, 2020.

[3] S. Kamiya, J.-K. Kang, J. Kim, A. Milidonis, and R. M. Stulz, "Risk management, firm reputation, and the impact of successful cyberattacks on target firms," *Journal of Financial Economics*, vol. 139, no. 3, pp. 719–749, 2021.

[4] M. S. K. Awan and L. Dahabiyeh, "Corporate attractiveness index: A measure for assessing the potential of a cyber attack," in *2018 9th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2018, pp. 1–6.

[5] G. Rafaiani, M. Battaglioni, S. Compagnoni, L. Senigagliesi, F. Chiaraluce, and M. Baldi, "A machine learning-based method for cyber risk assessment," in *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2023, pp. 263–268.

[6] M. F. Franco, E. Sula, A. Huertas, E. J. Scheid, L. Z. Granville, and B. Stiller, "Secriskai: A machine learning-based approach for cybersecurity risk prediction in businesses," in *2022 IEEE 24th Conference on Business Informatics (CBI)*, vol. 1. IEEE, 2022, pp. 1–10.

[7] F. Caron, J. Vanthienen, and B. Baesens, "Rule-based business process mining: Applications for management," in *Management Intelligent Systems: First International Symposium*. Springer, 2012, pp. 273–282.

[8] W. Cai and H. Yao, "Research on information security risk assessment method based on fuzzy rule set," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 9663520, 2021.

[9] P. Lou, G. Lu, X. Jiang, Z. Xiao, J. Hu, and J. Yan, "Cyber intrusion detection through association rule mining on multi-source logs," *Applied Intelligence*, vol. 51, pp. 4043–4057, 2021.

[10] F. Caron, J. Vanthienen, and B. Baesens, "Comprehensive rule-based compliance checking and risk management with process mining," *Decision Support Systems*, vol. 54, no. 3, pp. 1357–1369, 2013.

[11] M. Lee, J. Choi, C. Choi, and P. Kim, "Apt attack behavior pattern mining using the fp-growth algorithm," in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2017, pp. 1–4.

[12] S. Szymoniak and K. Foks, "Open source intelligence opportunities and challenges–a review," *Advances in Science and Technology. Research Journal*, vol. 18, no. 3, 2024.

[13] R. Jamilov, H. Rey, and A. Tahoun, "The anatomy of cyber risk," National Bureau of Economic Research, Tech. Rep., 2021.

[14] P. Giudici and E. Raffinetti, "Explainable ai methods in cyber risk management," *Quality and reliability engineering international*, vol. 38, no. 3, pp. 1318–1326, 2022.

The Dark Side of Digital Twins: Adversarial Attacks on AI-Driven Water Forecasting

Mohammadhossein Homaei
Grupo de Ingeniería de Medios
Universidad de Extremadura
Mhomaein@alumnos.unex.es

Víctor González Morales,
Óscar Mogollón-Gutiérrez
Grupo de Ingeniería de Medios
Universidad de Extremadura
{victorgomo, oscarmg}@unex.es

Andrés Caro
Grupo de Ingeniería de Medios
Universidad de Extremadura
andresc@unex.es

***Abstract*—Digital twins (DTs) are improving water distribution systems by using real-time data, analytics, and prediction models to optimize operations. This paper presents a DT platform designed for a Spanish water supply network, utilizing Long Short-Term Memory (LSTM) networks to predict water consumption. However, machine learning models are vulnerable to adversarial attacks, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These attacks manipulate critical model parameters, injecting subtle distortions that degrade forecasting accuracy. To further exploit these vulnerabilities, we introduce a Learning Automata (LA) and Random LA-based approach that dynamically adjusts perturbations, making adversarial attacks more difficult to detect. Experimental results show that this approach significantly impacts prediction reliability, causing the Mean Absolute Percentage Error (MAPE) to rise from 26% to over 35%. Moreover, adaptive attack strategies amplify this effect, highlighting cybersecurity risks in AI-driven DTs. These findings emphasize the urgent need for robust defenses, including adversarial training, anomaly detection, and secure data pipelines.**

***Index Terms*—Digital Twins, Artificial Intelligence, Cybersecurity, Adversarial Machine Learning Attack**

I. Introduction

Digital twin (DT) technology is helping many industries by improving how they simulate and manage physical systems. In water distribution networks, DTs allow real-time monitoring and forecasting, making operations more efficient and sustainable [1], [2].

However, DTs also face serious cybersecurity risks due to their complexity and constant internet connection. One major threat is adversarial machine learning (AML), where attackers manipulate input data (data poisoning) or the model itself (model poisoning) [3], [4]. This can reduce forecast accuracy and increase system costs.

Attacks like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) can create small changes in input data that make models like LSTM give wrong predictions [5]–[8]. Since LSTM models are widely used for water consumption forecasting [9], [10], these attacks can damage system performance.

This paper focuses on cybersecurity risks in DTs for water distribution networks, with a focus on forecasting water consumption using LSTM models. While DTs help improve efficiency, their connection to IoT and AI systems makes them vulnerable to AML attacks. Methods like FGSM and PGD can reduce prediction accuracy and raise costs by adding small changes to input data. To make these attacks harder to detect, this study explores the use of LA to adjust the perturbations. Additionally, it also proposes defense strategies to protect DTs from data and model poisoning, improving system reliability and supporting safer water management.

This paper is structured as follows: Section II reviews related work and the motivation behind addressing cybersecurity in WDS. Section III presents the proposed DT platform and forecasting models. Section IV analyzes the impact of FGSM-based AML attacks on LSTM models. Section V applies the FGSM attack and evaluates forecasting accuracy. Section VI introduces a Learning Automata-based FGSM attack to improve stealth. Section VII proposes a Random Learning Automata strategy to enhance unpredictability. Section VIII discusses DT vulnerabilities and mitigation strategies. Finally, Section IX concludes the paper and outlines future directions.

II. Related Works

The increasing use of AI and ML in water distribution networks has led to important progress in digital twin (DT) technology. However, these AI-powered systems also bring cybersecurity problems, especially adversarial machine learning (AML) threats. Several studies have explored the combination of DTs, AI, and cybersecurity in water management.

One important research area is the use of AI models like LSTM in water forecasting. Works such as [9], [10] have shown that LSTM models can predict water consumption with good accuracy. But these models are still vulnerable to adversarial attacks that reduce prediction quality.

Another line of research studies how to protect AI-based water systems from cyberattacks. For example, [6], [7] examined FGSM attacks on AI infrastructure, showing that even small changes in input data can cause big errors, highlighting the importance of strong security strategies.

Additionally, several DT projects in the water sector have used advanced ML to improve monitoring. [11] integrated graph convolutional networks (GCNs) in DTs for hydraulic state estimation. Similarly, [12] used graph neural networks (GNNs) for better sensor placement.

Despite these advancements, limited research has focused on the impact of AML threats on digital twin systems in the water industry. Most works study general cybersecurity or ML problems, but not how adversarial attacks reduce the accuracy of water forecasting. This gap shows the need for more research to protect AI-based DTs from these specific threats.

Table I
DT PROJECTS IN WATER INDUSTRY WITH AI/ML/DL AND THEIR VULNERABILITIES TO POISONING ATTACKS

Project	Data Used	Model/Algorithm	ML/AI/DL Techniques	Vulnerability	Possible Attack Vectors	Mitigation Strategies
Ciliberti et al. (2021)	Pressure, flow, and asset data	AI models for leak detection, optimization	ML/AI for DMA optimization	High	Data poisoning of asset management systems	Data integrity verification, encrypted data channels
Bonilla et al. (2022)	Real-time pressure and flow rate data	GCNs	AI/ML for hydraulic state estimation	High	Data poisoning of real-time pressure and flow data	Data validation techniques, secure data pipelines
Zekri et al. (2022)	IoT sensor data, asset operation data	Multi-Agent Systems, AI-driven agents	Multi-agent reinforcement learning	Moderate-High	Model poisoning by corrupting reward system	Robust reward functions, anomaly detection
Matheri et al. (2022)	Wastewater treatment sensor data	Cyber-Physical Systems (CPS), AI optimization models	AI/ML for predictive maintenance	High	Data poisoning via corrupted sensor data	Secure real-time data transmission, predictive anomaly detection
Ramos et al. (2022)	Leakage detection, water usage data	Optimization algorithms	AI-driven optimization for water management	Moderate	Tampered leakage data poisoning	Cryptographic methods for data validation
Henriksen et al. (2022)	Hydrological data, climate models	ML Models	ML for climate adaptation	Moderate	Tampered hydrological data affecting water management	Redundant data sources, data quality audits
Savic (2022)	Water usage patterns	AI anomaly detection models	AI/ML for anomaly detection	High	Model poisoning via corrupted anomaly detection data	AI model validation, adversarial training
Pedersen et al. (2022)	Water level sensors, drainage data	Hydraulic models with DTs	ML for error classification	Moderate	Data poisoning from water level sensors	Redundant sensor validation, secure data transmission
Valencia Smart Water (2023)	SCADA, customer feedback	Hydraulic models with real-time optimization	AI/ML for pressure management	Moderate	Tampered sensor inputs poisoning	Secure data transmission, blockchain verification
Sabesp Digital Twin (2023)	IoT sensor data, remote monitoring	ML for anomaly detection	AI/ML for fault detection	High	Data poisoning from faulty sensors or pump data manipulation	Real-time anomaly detection, secure IoT devices
Tarragona Water Consortium (2023)	Hydraulic models, real-time sensors	Live simulations for predictive maintenance	ML for predictive maintenance	Moderate	Data manipulation through compromised sensors	Redundant sensor data validation systems
Smart Water Grid in Gaula (2023)	Leakage detection, water usage data	Digital twin with real-time optimization	AI-driven optimization for water loss prevention	Moderate-High	Data poisoning from compromised sensors	Blockchain-based validation, real-time monitoring
Water Research Foundation AI/ML Project (2023)	Utility performance data	ML models for prediction	AI/ML for performance optimization	High	Model poisoning via contaminated datasets	Regular model audits, federated learning
Menapace et al. (2024)	Pressure sensor data	GNNs	DL/ML for pressure estimation	High	Data poisoning through compromised sensor data	Data cross-validation, hybrid training with anomaly detection

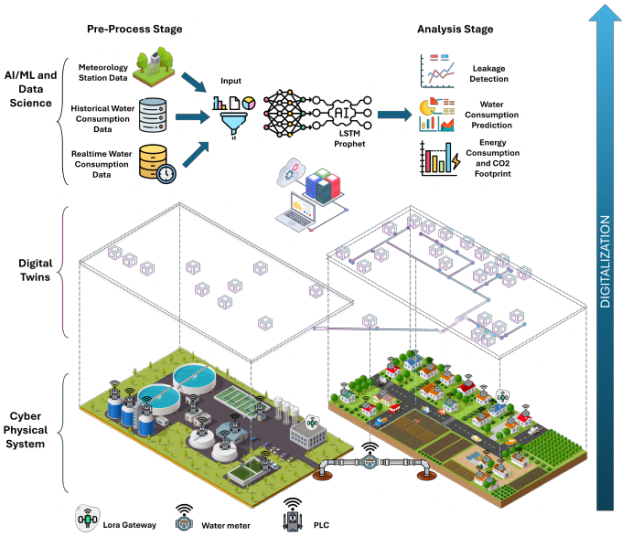


Figure 1. DT Platform in the Water Distribution Networks

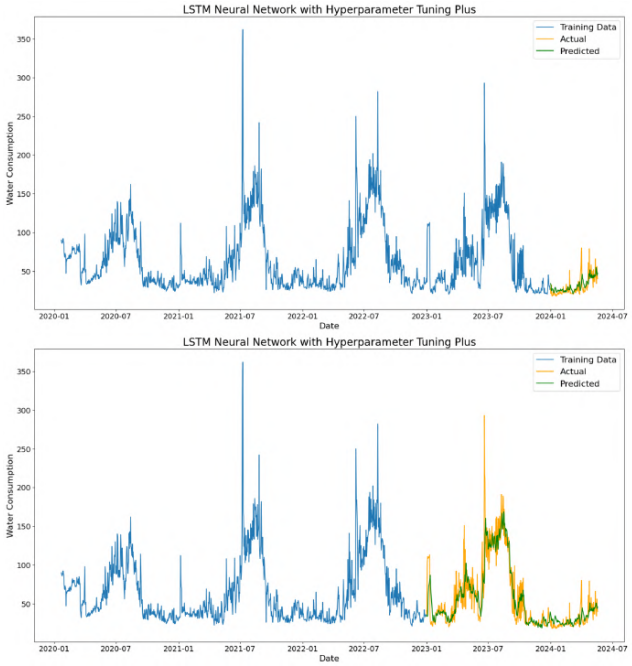


Figure 2. Water consumption forecasting via LSTM for 6(top) and 18 months(bottom)

III. PROPOSED DT AND FORECASTING MODELS

CAUCCES is a DT platform designed to enhance water distribution through real-time monitoring, predictive analysis, and data-driven decision-making (showed in Figure 1). It has been developed in collaboration with the Media Engineering Group at the University of Extremadura and Ambling Ingeniería y Servicios, S.L [13]. The platform integrates IoT sensors, AI-based forecasting, and secure data management to improve efficiency and minimize water loss.

Traditional water networks lack smart monitoring and forecasting, making them vulnerable to challenges such as aging infrastructure and environmental impacts. CAUCCES addresses these issues by continuously gathering data, utilizing reliable communication technologies, and optimizing scheduling for better maintenance and operation. This creates a real-time digital replica of the water system, enabling early detection and prevention of potential problems before they affect service. Figure 2 presents the forecasting results of water consumption using the UV-LSTM models.

IV. AML ATTACKS ON FORECASTING

AML is an emerging field within the broader domain of ML, focusing on designing and evaluating models that are vulnerable to adversarial inputs. These inputs, also called adversarial examples, are carefully crafted perturbations designed to mislead the model into making incorrect predictions while appearing normal to humans. Such perturbations are typically small enough to go unnoticed in the input data but large enough to degrade the model's performance significantly. AML is crucial in various applications, particularly in domains where high reliability is required, such as healthcare, autonomous systems, and financial forecasting.

One basic AML method is the FGSM. It is a white-box attack, meaning the attacker knows the model's details. FGSM uses the gradient of the loss function to add small changes to the input that increase prediction error. The formula is:

$$X_{\text{adv}} = X + \epsilon \cdot \text{sign}(\nabla_X J(\theta, X, y)) \tag{1}$$

Where:

- X_{adv} is the adversarial input generated by the FGSM attack.
- X is the original input data (e.g., daily temperature and water consumption records).
- ϵ is the perturbation magnitude, determining the intensity of the attack.
- $J(\theta, X, y)$ represents the loss function, with θ denoting the model parameters and y being the true output label.
- $\nabla_X J(\theta, X, y)$ is the gradient of the loss function for the input data.

Meanwhile, the core formula for generating adversarial examples using PGD is as follows:

$$X_{\text{adv}}^{(t+1)} = \text{clip}_{[X-\epsilon,\, X+\epsilon]}\Big(X_{\text{adv}}^{(t)} + \alpha \cdot \text{sign}\big(\nabla_{X_{\text{adv}}^{t}} J\big(X_{\text{adv}}^{(t)}, y\big)\big)\Big) \tag{2}$$

Where:

- $X_{\text{adv}}^{(t)}$ refers to the generation of adversarial attacks at the time step t in the iterative sequence of the generation of FGSM attacks. At $t = 0$, it would be the input (daily water consumption or temperature).
- α is the step size at each iteration.
- $\text{sign}(\cdot)$ returns the sign of each component.
- The clip function projects the example into the ϵ-ball around x and enforces the valid data range.

PGD is more effective than FGSM because it applies small steps iteratively. Both methods are commonly used to test model robustness against adversarial inputs.

V. Applying FGSM to the LSTM Model for Water Consumption Forecasting

The LSTM model used for forecasting water consumption from daily usage and temperature data can be tested against FGSM adversarial attacks. Since LSTM handles time series, small input changes may spread over time and cause large prediction errors.

To apply FGSM, the first step is to compute the gradient of the loss function with respect to the input sequence. The goal is to slightly modify each feature (temperature and consumption) at every time step, creating a perturbed input X_{adv} that makes the model's prediction less accurate.

The adversarial input is then passed to the LSTM to measure how much its performance changes. As shown in Table II, even a small perturbation (ϵ=0.001) increases MAE and RMSE slightly. When ϵ is larger (e.g., 0.01), errors grow more, with MAE over 18 and MAPE above 35%. This shows the LSTM model is sensitive to small adversarial changes in its input.

To further evaluate the model's resilience, PGD is applied. Unlike FGSM, PGD refines the perturbation iteratively, leading to a stronger adversarial impact. Table III shows that for ϵ=0.01, the MAE increases at a similar rate as FGSM, but for lower values of ϵ (e.g., 0.005 and 0.008), the prediction error already exhibits a steeper increase in RMSE and MAPE compared to FGSM. This suggests that even at intermediate perturbation levels, PGD induces more severe deviations in the LSTM model's forecasts.

Table II
Results of FGSM attack with different ϵ values on LSTM models

Model	ϵ	MAE	RMSE	MAPE
LSTM	0.0	12.334	20.653	24.419
LSTM	0.001	12.660	20.861	25.261
LSTM	0.005	13.959	21.718	28.609
LSTM	0.008	14.926	22.385	31.104
LSTM	0.01	15.567	22.840	32.754
LSTM+	0.0	12.329	20.449	25.032
LSTM+	0.001	12.758	20.726	26.108
LSTM+	0.005	14.462	21.879	30.378
LSTM+	0.008	15.721	22.783	33.531
LSTM+	0.01	16.551	23.400	35.606

Table III
Results of PGD attack with different epsilon values on LSTM models

Model	ϵ	MAE	RMSE	MAPE
LSTM	0.0	12.334	20.653	24.419
LSTM	0.001	12.660	20.861	25.261
LSTM	0.005	13.960	21.720	28.615
LSTM	0.008	14.933	22.390	31.125
LSTM	0.01	15.579	22.848	32.790
LSTM+	0.0	12.329	20.449	25.032
LSTM+	0.001	12.759	20.726	26.108
LSTM+	0.005	14.467	21.884	30.395
LSTM+	0.008	15.741	22.797	33.588
LSTM+	0.01	16.589	23.424	35.708

Comparing both attacks, PGD consistently leads to higher errors at every tested epsilon value. While FGSM causes a steady degradation in model accuracy, PGD intensifies this effect by iteratively optimizing the perturbation, making it more effective in misleading the model. Notably, for ϵ=0.005, the difference between FGSM and PGD is already evident in all error metrics, particularly in RMSE.

VI. LA-Based Undetectable FGSM Attack

This section presents a learning automata-based approach for dynamically adjusting the perturbation size in the FGSM attack on LSTM models [14]. The primary objective is to improve attack stealth while maintaining its effectiveness in reducing forecasting accuracy.

Using learning automata, the ϵ values were adjusted iteratively to maintain an effective yet undetectable attack. The MAPE progression over iterations showed a smooth increase, avoiding sharp variations, as shown in 3. The probability evolution of different ϵ values demonstrated the learning automata's ability to converge toward optimal attack parameters.

Figure 4 illustrates how an FGSM attack based on LA can remain imperceptible to a human observer. LA introduces

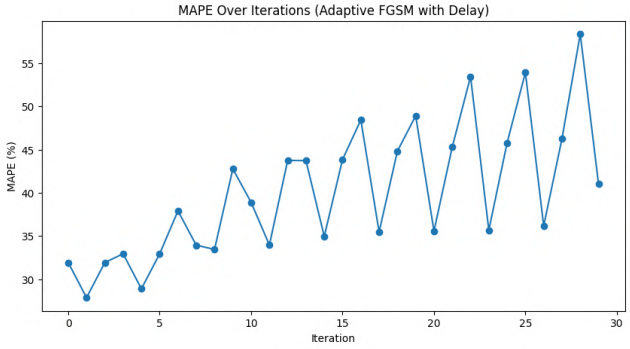


Figure 3. Fluctuation of the epsilon variable along the iterations in an LA-based FGSM Attack

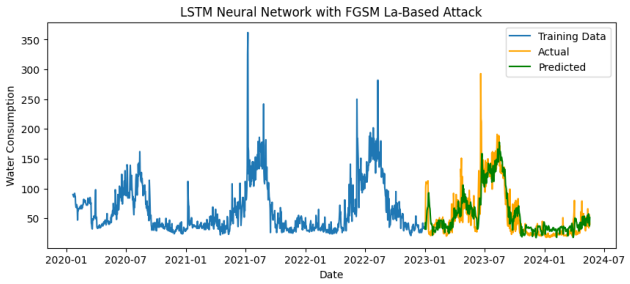


Figure 4. Hidden LA-based FGSM Attack

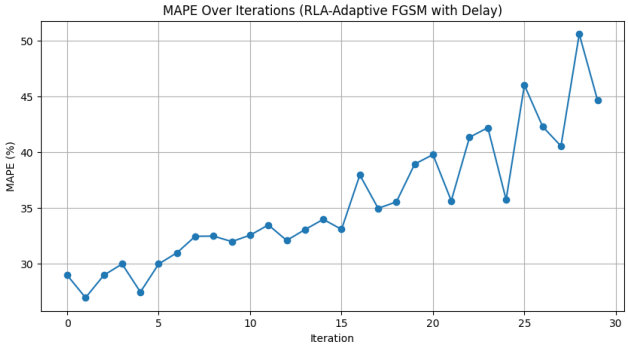


Figure 5. Fluctuation of the epsilon variable along the iterations in a RLA-based FGSM Attack

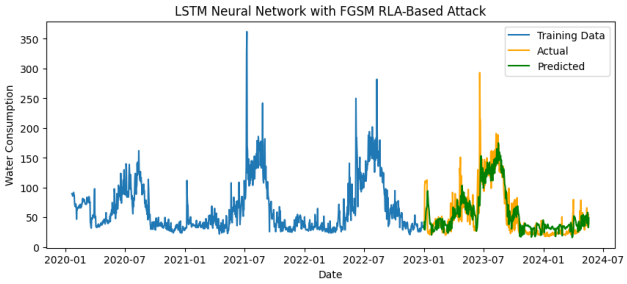


Figure 6. Hidden RLA-based FGSM Attack

variability, preventing a straightforward detection pattern. This alternation in perturbation intensity adds a layer of randomness that disrupts the usual correlation between distortion and detectability. As a result, the attack does not exhibit a consistently increasing trend, making it more challenging to distinguish from natural fluctuations in the data. This adaptive behavior enhances the stealthiness of the adversarial perturbations, posing greater difficulties for both manual and automated detection methods.

VII. Random Learning Automata for FGSM Attack Optimization

This section presents a Random Learning Automata (RLA)-based approach for dynamically adjusting the perturbation size in the FGSM attack on LSTM models. The key idea is to select multiple ϵ values per iteration instead of a single value, which helps improve attack stealth and avoid detection by the forecasting model [14].

In contrast to standard learning automata, RLA selects a random combination of ϵ values from a predefined set instead of a single value.

At each iteration, a subset of ϵ values is selected with probabilities that will be adjusted after each iteration.

To ensure the adaptive selection of ϵ values, the probability update mechanism follows:

$$P(a_t) = P(a_t) + r \cdot (1 - P(a_t)), \quad \text{if rewarded} \tag{3}$$

$$P(a_t) = P(a_t) \cdot (1 - p), \quad \text{if penalized} \tag{4}$$

Instead of choosing only one ϵ per iteration, RLA selects two or three ϵ values and applies them simultaneously.

The attack's effectiveness is measured using the MAPE metric. To prevent abrupt changes in the attack, the penalty factor is adjusted dynamically:

$$p_{\text{adaptive}} = \begin{cases} 3p, & \text{if MAPE} > 100\% \\ 1.5p, & \text{if } \Delta\text{MAPE} > 5\% \\ p, & \text{otherwise} \end{cases} \tag{5}$$

To further enhance stealth, adversarial examples are stored and applied after a delay. Instead of using the perturbed input immediately, RLA waits for a iterations before introducing the modified input into the model.

A. Experimental Results

The FGSM attack using RLA adds more unpredictability to the perturbation process. Unlike fixed learning methods, RLA selects ϵ values randomly, using reward and penalty rules, as shown in Figure 5. This randomness avoids clear patterns, making the attack look chaotic and harder to detect with common anomaly methods.

A key point of RLA is that it does not follow a fixed rule. Instead, ϵ values change in a random way, similar to chaotic systems. This makes it difficult to identify repeated patterns, which complicates both manual and automatic defenses. The randomness helps the attack stay hidden and avoid detection.

This effect can be seen in how ϵ values change during the attack. Because of this variation, it is difficult to clearly separate normal data from adversarial data. Figure 6 shows how the changing ϵ values make the attack more stealthy.

Table IV
MITIGATION STRATEGIES FOR ADVERSARIAL ATTACKS IN DIGITAL TWIN FORECASTING MODELS

Aspect	Data / System	Technique / Method	Potential Vulnerability	Risk Level	Possible Attack Vectors	Mitigation Strategies
LoRa Encryption	Meter data, device provisioning	Built-in AES-128	Poor key management, reuse of keys	Moderate	Key guessing, eavesdropping on packets	Frequent key rotation, secure key distribution, proper implementation of AES-128
Meter-to-Gateway Sec	LoRa meter to LoRa gateway link	Mutual authentication, secure join procedures	Replay or spoofing of meter credentials	High	Impersonation of valid meters, data injection	Challenge-response protocols, nonce usage, short-lived session keys
Firmware Integrity	On-device firmware (water meters, gateways)	Signed or hashed firmware updates	Unauthorized firmware modifications	High	Malicious updates, remote code execution	Secure OTA updates with signature checks; regular patching
Gateway-to-Server	Data in transit from gateway to server	TLS/SSL, VPN, or private lines	Man-in-the-middle attacks	Moderate	Traffic interception, unauthorized data reading	Encrypted communication tunnels, certificate-based authentication
Net Monitoring & IDS	LoRa gateway and backend network	Intrusion Detection / Prevention Systems	Undetected brute-force, scanning attempts	Moderate	Malicious traffic patterns, repeated authentication failures	Automated anomaly detection, real-time threat response
ChirpStack Security	ChirpStack network server, application server	Role-based access, secure APIs	Misconfiguration, weak API keys	Moderate	Unauthorized device provisioning, data leakage	Secure API endpoints, regular security audits, minimal privilege policies
Database Security (PostgreSQL)	Stored water consumption records	Encryption at rest, access controls	Unauthorized DB access or tampering	High	SQL injection, stolen credentials	Strict role management, periodic audits, row-level security
End-to-End Encryption	Full data flow from meter to final storage	Consistent encryption in transit and at rest	Partial encryption gaps	Moderate	Plaintext exposure at intermediate hops, data sniffing	Holistic encryption approach, verifying encryption at all layers
AML Training for AI	Forecasting model (e.g., LSTM)	Incorporating FGSM/PGD examples into training	Model easily fooled by small perturbations	High	Data poisoning, gradient-based adversarial attacks	Model retraining on adversarial samples, ensemble methods
Domain-based Constraints	Forecasts	Physical/hydraulic plausibility checks	Acceptance of impossible meter readings	Moderate	Sudden large outliers can corrupt predictions	Filter or flag data outside valid usage/pressure thresholds
Real-time Anomaly Detection	Incoming meter data stream	Isolation Forest, One-Class SVM	Persistent adversarial or sensor tampering	High	Silent data drift, gradual poisoning	Trigger alerts when usage deviates from historical/seasonal norms
Gradient Masking & Model Randomization	Neural network layers	Adding noise, dropout, random inference steps	Straightforward gradient-based attack	Moderate	White-box adversary calculates precise gradients	Stochastic layers obscure gradients, raising attack complexity
Key Management & Regular Audits	All cryptographic operations	Rotating keys, HSM usage, compliance checks	Stolen or expired keys, unpatched systems	Moderate	Privilege escalation, extended infiltration	Automated key rotation, routine compliance (ISO/IEC 27001), robust backup/redundancy

VIII. MITIGATION STRATEGIES

Protecting digital twin (DT) systems from adversarial attacks requires a combination of cybersecurity measures, data integrity techniques, and machine learning defenses (Table IV). Key strategies focus on strengthening AI models against manipulation, ensuring secure data transmission, and implementing real-time monitoring. AI-based anomaly detection can identify suspicious activity, while adversarial training improves model robustness. Secure encryption, authentication protocols, and strict access controls help prevent unauthorized access and data tampering. Additionally, continuous auditing and compliance with cybersecurity standards ensure long-term resilience. By integrating these strategies, DTs can maintain reliable forecasting and decision-making even in the presence of cyber threats.

ACKNOWLEDGEMENT

This initiative is carried out within the framework of the funds from the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation) – National Institute of Cybersecurity (INCIBE), as part of project C107/23: "Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures."

IX. CONCLUSION AND FUTURE WORK

This study shows that although AI-based DTs are helpful for water forecasting and resource management, they can still be attacked. Small, hidden changes in the data—called adversarial attacks—can reduce LSTM accuracy, increase costs, and damage trust. Our research highlights this hidden danger, which is important for city infrastructure and water systems where wrong decisions can have big impacts. Our method using learning automata can adapt to and follow the monthly and seasonal fluctuations in water consumption patterns.

Next, we plan to use Zabbix for real-time monitoring of things like sensor status, unusual forecasts, and network traffic. With smart alert settings, Zabbix can find problems fast and react automatically. In future work, we also want to use federated learning to avoid having one weak point and explore using multiple models together for stronger defense. These steps will help protect DTs and keep water systems safer from cyber threats.

REFERENCES

[1] A.-J. Wang, H. Li, Z. He, Y. Tao, H. Wang, M. Yang, D. Savic, G. T. Daigger, and N. Ren, "Digital twins for wastewater treatment: A technical review," *Engineering*, vol. 36, p. 21–35, May 2024.

[2] H. Beji and M. Lade, "Impact of digital transformation on carbon emissions reductions in the water industry," in *Lecture Notes in Energy*. Springer International Publishing, 2022, pp. 117–127.

[3] M. H. Homaei, A. C. Lindo, J. C. S. Núñez, O. M. Gutiérrez, and J. A. Díaz, Eds., *The role of Artificial Intelligence in Digital Twin's Cybersecurity*. Editorial Universidad de Cantabria, Sep. 2022.

[4] M. Homaei, Óscar Mogollón-Gutiérrez, J. C. Sancho, M. Ávila, and A. Caro, "A review of digital twins and their application in cybersecurity based on artificial intelligence," *Artificial Intelligence Review*, vol. 57, no. 8, p. 201, Jul. 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10805-3

[5] E. Coda, B. Clymer, C. DeSmet, Y. Watkins, and M. Girard, "Universal fourier attack for time series," *IEEE Open Journal of Signal Processing*, vol. 5, p. 858–866, 2024.

[6] S. U. Khan, M. Mynuddin, and M. Nabil, "Adaptedge: Targeted universal adversarial attacks on time series data in smart grids," *IEEE Transactions on Smart Grid*, vol. 15, no. 5, p. 5072–5086, Sep. 2024.

[7] M. Mynuddin, S. U. Khan, R. Ahmari, L. Landivar, M. N. Mahmoud, and A. Homaifar, "Trojan attack and defense for deep learning-based navigation systems of unmanned aerial vehicles," *IEEE Access*, vol. 12, p. 89887–89907, 2024. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2024.3419800

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019. [Online]. Available: https://arxiv.org/abs/1706.06083

[9] A. Niknam, H. K. Zare, H. Hosseininasab, and A. Mostafaeipour, "Developing an lstm model to forecast the monthly water consumption according to the effects of the climatic factors in yazd, iran," *Journal of Engineering Research*, vol. 11, no. 1, p. 100028, Mar. 2023.

[10] R. Qiu, Y. Wang, B. Rhoads, D. Wang, W. Qiu, Y. Tao, and J. Wu, "River water temperature forecasting using a deep learning method," *Journal of Hydrology*, vol. 595, p. 126016, Apr. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.jhydrol.2021.126016

[11] C. A. Bonilla, A. Zanfei, B. Brentan, I. Montalvo, and J. Izquierdo, "A digital twin of a water distribution system by using graph convolutional networks for pump speed-based state estimation," *Water*, vol. 14, no. 4, p. 514, Feb. 2022. [Online]. Available: http://dx.doi.org/10.3390/w14040514

[12] A. Menapace, A. Zanfei, M. Herrera, and B. Brentan, "Graph neural networks for sensor placement: A proof of concept towards a digital twin of water distribution systems," *Water*, vol. 16, no. 13, p. 1835, Jun. 2024.

[13] Ambling Ingeniería y Servicios, S.L., "Ambling. Official Website," 2025, accessed: March 20, 2025. [Online]. Available: https://www.ambling.es

[14] A. Rezvanian, A. M. Saghiri, S. M. Vahidipour, M. Esnaashari, and M. R. Meybodi, *Recent Advances in Learning Automata*. Springer International Publishing, 2018.

A Review of: Protocol Dialects as Formal Patterns

Daniel Galán
ETH Zurich
Zürich, Switzerland
daniel.galan@inf.ethz.ch

Víctor García
VRAIN, Universitat Politècnica de València
Valencia, Spain
vicgarval@upv.es

Santiago Escobar
VRAIN, Universitat Politècnica de València
Valencia, Spain
sescobar@upv.es

Catherine Meadows
Naval Research Laboratory
Washington DC, USA
catherine.meadows@nrl.navy.mil

Jose Meseguer
University of Illinois at Urbana-Champaign
Illinois, USA
meseguer@illinois.edu

***Abstract*—A *protocol dialect* is a lightweight method to obfuscate the communication exchanges between legitimate protocol users to make it hard for malicious users to interact with legitimate ones. So far, dialects have been based on a single obfuscation transformation, which we call a *lingo*. In this work dialects are generalized to become *protocol and lingo generic*. In this way they can be *composed* with other dialects for greater security. We present a formal semantics of dialects as *formal patterns*, i.e., as protocol transformations formalized in rewriting logic. We also present several *attacker models* and explain how our generalized notion of dialect can be used to harden dialects against such attackers.**

***Index Terms*—Protocols, Dialects, Formal Patterns, Moving Target Defense, Security**

Contribution type: *Research already published*

I. Introduction

Protocol dialects are modifications to existing protocols that provide lightweight security, especially against relatively easy attacks that could potentially be leveraged into more serious ones. The scenario is that of a network of mutually trusting principals, e.g. an enterprise network, that needs to defend itself from outside attackers. The most effective way is to prevent a potential attacker from even initiating a communication with group members, e.g. by requiring messages to be modified in some unpredictable way by means of a *dialect-based* modification of the protocol messages, that cannot be easily reproduced by the attacker.

In this paper, already published in [1], we develop a framework that provides a formal basis for reasoning about dialects and their properties. Our framework is based on *formal patterns*, which allow us to generalize the notion of dialect so as to make it both protocol- and lingo- generic. This allows not only arbitrary classes of lingos, but even arbitrary combinations of lingos. This genericity has two important consequences: (1) dialects now become *composable*; and (2) composed dialects can be much harder to break than the simpler dialects they are composed out of.

II. Dialects as Formal Patterns

We generalize the notion of *dialect*, which itself uses the notion of a *lingo*, i.e., an invertible message transformation. We make dialects *generic* on two dimensions: (i) *protocol generic*, so as to apply to a broad class of protocols *without changing the code of the original protocol*; and (ii) *lingo generic*, so that a dialect can, on the one hand, be instantiated with different lingos, and, on the other hand, may use *several lingos simultaneously* to achieve higher levels of obfuscation.

Fig. 1 describes our approach. It depicts two communicating objects of $\mathcal{P}$ (in purple) wrapped inside their dialect meta-objects (in blue). Messages are depicted as envelopes. The objects of $\mathcal{P}$ behave exactly as before: their meta-objects take care of obfuscating and de-obfuscating their messages. The lingo currently used takes the form of a pair of (parametric) functions (f, g), where f obfuscates messages and g puts them back in the format expected by participants in the original protocol $\mathcal{P}$. Our approach has two important advantages. First, it applies to unlimited classes of protocols without requiring any modifications in their code. Second, by using several lingos simultaneously, it can make the task of a malicious intruder much harder.

A. Lingos

As described in Fig. 1, dialect wrappers use invertible *message transformations* called *lingos* (denoted f and g in the picture) to obfuscate a message m while in transit. These transformations are in fact *parametric*, with the parameter, say, a, ranging over some data type A. Lingos are also *theory-generic*, that is, they apply to and *extend* not just a single actor rewrite theory but a class of so-called *protocol theories*. Note that they are *parametrically inverse functions*, i.e., $g(f(\vec{m}, a), a) = \vec{m}$.

B. Dialects

Dialects are protocol transformations of the form $\mathcal{P} \mapsto \mathcal{D}(\mathcal{P}, \Lambda_1(\mathcal{P}), \ldots, \Lambda_N(\mathcal{P}))$, where $\mathcal{D}(\mathcal{P}, \Lambda_1(\mathcal{P}), \ldots, \Lambda_N(\mathcal{P}))$ endows $\mathcal{P}$ with N lingos, and wraps each protocol participant inside a *dialect meta-actor* that uses lingos $\Lambda_1(\mathcal{P}), \ldots, \Lambda_N(\mathcal{P})$ to obfuscate the communication between the honest protocol actors of $\mathcal{P}$.

III. Dialects as Moving Targets

Every dialect should be a moving target. But some dialects move faster than others. A *static* dialect is one where the *lingos* of a dialect never change. It is still a moving target, because the *parameters* of the fixed lingo or lingos used are always changing. Faster moving targets can be achieved by means of what we call *dynamic* dialects, where the *lingos* of the dialect change dynamically. This frequent change of the communication format imposes a serious burden on intruders,

ISBN:978-84-10169-61-6

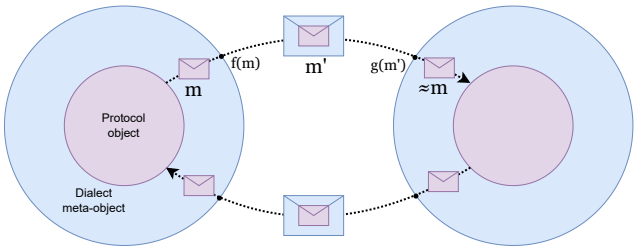


Fig. 1. Representation of dialects as meta-objects and their interactions.

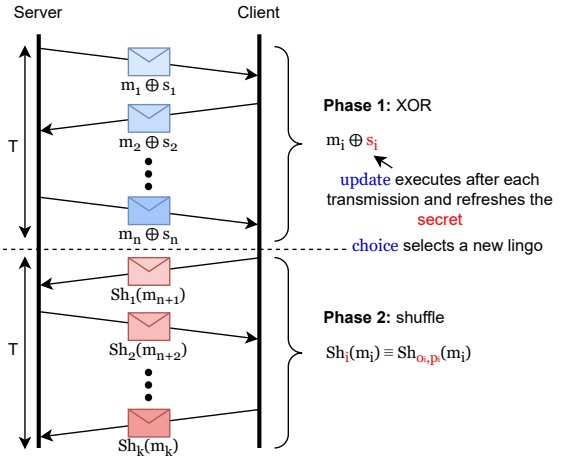


Fig. 2. Example sequence of two principals using a dynamic periodic dialect.

because even if a specific lingo is successfully broken, the dialect will change this lingo after some time, forcing the intruder to start all over again its malicious efforts.

A. Periodic and Aperiodic Dialects

We distinguish two kinds of dialects according to how the new lingo is selected:

(i) **Periodic** dialects include Time and change their current lingos every time a predefined period T elapses. More complex periodic dialects can be defined, where the period keeps changing as time phases advance.

(ii) **Aperiodic** dialects change their current lingos depending on the history of previous messages rather than depending on time.

We can illustrate the behavior of a periodic dialect with the example in Fig. 2. The vertical axis represents time advancing from top to bottom. During a *time phase*, all participants of a particular modified protocol share the same set of *current lingos*. At *phase boundaries*, the phase changes, and then dialects use their choice function to agree on a new set of current lingos, as depicted by the dashed horizontal line. In the figure, during the first phase `XOR` is the current lingo, and `shuffle` becomes the current lingo during the second phase. Note that during each phase parameters $s_1, s_2, \ldots$ for `XOR`, resp. $\{o_1, p_1\}, \{o_2, p_2\}, \ldots$ for `shuffle`, change all the time in a synchronized fashion between server and clients.

IV. Dialect Composition Operations

Our formal framework provides *dialect composition operations* to generate new, more sophisticated dialects out of simpler ones.

There are two degrees of freedom enabling a dialect to become more complex: (i) increasing the complexity of the lingos, and (ii) increasing the number of lingos available. We then distinguish two kinds of dialect composition: (a) *vertical* composition increases both degrees of freedom (i)–(ii); and (b) *horizontal* composition increases degree of freedom (ii).

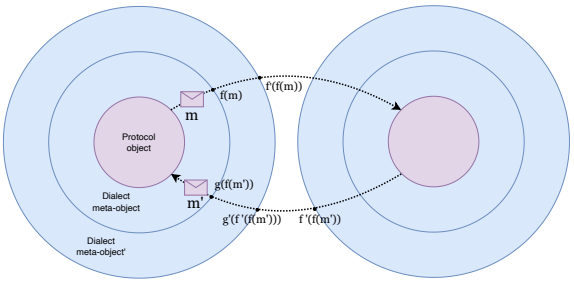


Fig. 3. Meta-object representation of the vertical composition of two dialects.

A. Vertical Composition

The vertical composition of two dialects combines the transformations expressed in their lingos in a function composition manner. From a meta-object point of view, vertical composition results in wrapping a second dialect meta-object on top of a first dialect meta-object that itself wraps an original protocol object, stacking them and thus further modifying the protocol messages.

As shown in Fig. 3, m is first transformed by f and then is further mutated by f', finally sending $f'(f(m))$ into the network. On the other side of the communication, the *incoming* message $f'(f(m))$ is reversed: first, g' is applied and then g. Of course, we can vertically compose not just two, but $n \geq 2$ dialects this way, thus resulting in a function pipeline $g(g'(g''(\ldots g'^n(f'^n(\ldots(f''(f'(f(m))))\ldots))\ldots)\ldots))$.

The dialects to compose must be chosen judiciously, as not all vertical compositions have the same properties.

B. Horizontal Composition

Horizontal composition takes the lingos chosen by two dialects, merges them together, and uses a pseudo-random function to choose a lingo in one of the two dialects. This operation can be iterated to horizontally compose $n \geq 2$ dialects.

When horizontally composing two dialects $\mathcal{D}$ and $\mathcal{D}'$, dialect designers can favor $\mathcal{D}$ over $\mathcal{D}'$ (or vice versa) by using the bias σ in the pseudo-random tossing a σ-biased coin to choose between $\mathcal{D}$ and $\mathcal{D}'$.

V. Conclusion

We have proposed modeling dialects as generic *formal patterns*, which transform protocols without changing their code. This is a vast generalization of the various dialects appeared so far in the literature that provides new methods to make dialects harder to break, e.g., by using *dynamic* dialects, and dialect *compositions*.

References

[1] D. Galán, V. García, S. Escobar, C. Meadows, and J. Meseguer, "Protocol dialects as formal patterns," in *European Symposium on Research in Computer Security*. Springer, 2023, pp. 42–61. [Online]. Available: https://doi.org/10.1007/978-3-031-51476-0_3

A review of: CYBERSHIELD: A Competitive Simulation Environment for Training AI in Cybersecurity

José Álvaro Fernández Carrasco, Iñigo Amonarriz Pagola, Raúl Orduna Urrutia, Rodrigo Román*
Vicomtech, Basque Research and Technology Alliance (BRTA), 20009 Donostia/San Sebastian, Spain
**Network Information and Computer Security Lab, University of Malaga, 29071 Malaga, Spain*
{*jafernandez, iamonarriz, rorduna*}*@vicomtech.org, rroman@uma.es*

***Resumen*—Los entornos de simulación son fundamentales en ciberseguridad para el entrenamiento de algoritmos de IA, proporcionando un espacio seguro y controlado donde analizar el impacto de diversas amenazas cibernéticas. Este artículo presenta CyberShield, un entorno avanzado de simulación diseñado para tareas de ciberseguridad, con un enfoque particular en la implementación y evaluación de algoritmos de IA, especialmente Reinforcement Learning. CyberShield simula una infraestructura con múltiples equipos, cada uno alojando distintos servicios con vulnerabilidades específicas. Dentro de este entorno, dos agentes opuestos (defensor y atacante) se enfrentan estratégicamente, cada uno con un conjunto de acciones diseñadas para superar al adversario. Optimizado para el entrenamiento competitivo de múltiples agentes mediante RL, CyberShield facilita la exploración de nuevas estrategias, promoviendo la automatización de tácticas ofensivas y defensivas.**

***Index Terms*—Reinforcement Learning, Respuesta, Markov, Moving Target Defense, Ciberseguridad, Multiagente**

Tipo de contribución: *Investigación ya publicada*

I. Introducción

Los entornos de simulación son fundamentales para el desarrollo seguro, eficiente y efectivo de algoritmos de IA en ciberseguridad. Proporcionan un espacio controlado donde los investigadores y desarrolladores pueden diseñar, evaluar y optimizar algoritmos, preparándolos para amenazas reales. En este contexto, ha aumentado el interés en la creación de entornos de simulación fiables para entrenar algoritmos de Reinforcement Learning (RL), con el objetivo de que los agentes entrenados se comporten de manera similar en escenarios reales. Este artículo presenta CyberShield [1], un entorno que simula la estructura de una red de comunicación compuesta por múltiples dispositivos interconectados, cada uno con vulnerabilidades y características específicas. El principal objetivo de este entorno es proporcionar un espacio controlado para la experimentación y evaluación de algoritmos de RL desde perspectivas tanto ofensivas como defensivas.

En el estado del arte actual, es poco común encontrar entornos que permitan entrenar y evaluar agentes de ataque y defensa de manera equilibrada. La mayoría de los entornos existentes están diseñados para optimizar o desarrollar estrategias centradas en un solo tipo de agente [2][3][4]. En cambio, CyberShield proporciona a cada agente un conjunto predefinido de acciones y observaciones del entorno, lo que permite a los desarrolladores entrenar agentes RL tanto ofensivos como defensivos y evaluar el rendimiento de sus algoritmos en un mismo marco de trabajo.

Además, CyberShield se plantea como una plataforma educativa al incluir un módulo interactivo en el que un usuario puede asumir el rol de uno de los agentes. Esta funcionalidad permite enfrentarse a distintos tipos de oponentes: un agente entrenado con RL, un agente que ejecuta acciones aleatorias o un agente basado en heurísticas.

II. Estado del Arte

El proceso de toma de decisiones en un sistema informático puede entenderse como un juego entre agentes adversarios: un atacante que busca comprometer la seguridad del sistema y un defensor que intenta neutralizar sus acciones y mantener el control [5][6]. Debido a esta dinámica, la comunidad científica trabaja activamente en herramientas de prueba que permitan analizar tácticas y estrategias desde perspectivas defensivas (blue agent) y ofensivas (red agent). Generalmente, los entornos se centran en tipos específicos de ataques con características predefinidas, priorizando la disrupción de redes (DDoS, fuzzing, jamming), la detección de intrusiones y el movimiento lateral dentro de la red para su expansión. La mayoría de los entornos existentes se centran en el desarrollo de uno de los dos agentes (defensor o atacante) [7][8][9][3]. Solo unos pocos han sido diseñados específicamente para el entrenamiento simultáneo de agentes atacantes y defensores, aunque presentan ciertas limitaciones. Por ejemplo, algunos entornos como [10] y [11] son altamente abstractos, con observaciones y acciones que no pueden traducirse directamente a un entorno de emulación realista. Otros, como [12], sí permiten el entrenamiento de agentes con acciones concretas, como reconocimiento o exploits para los atacantes, facilitando su despliegue en entornos de emulación. Sin embargo, presentan una limitación: aunque es posible entrenar agentes atacantes y defensores con RL, no pueden entrenarse simultáneamente, sino que es necesario seleccionar cuál de los dos será optimizado en cada sesión de entrenamiento.

III. Arquitectura de CyberShield

CyberShield modela una red de computadoras con múltiples servicios y vulnerabilidades (ver Figura 1). Su estructura permite la interacción entre un agente defensor y un atacante, cada uno con conjuntos de acciones predefinidos. Se basa en los principios de los Procesos de Decisión de Markov (MDP), lo que permite un entrenamiento estructurado de los agentes.

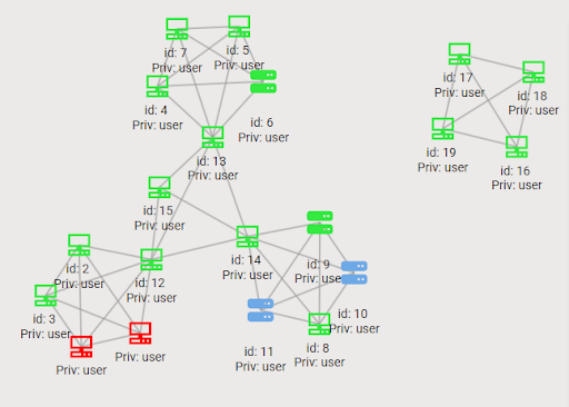


Figura 1. CyberShield

III-A. Espacio de Acciones

Cada agente tiene un conjunto específico de acciones disponibles, diseñadas para simular de manera realista los ataques y defensas en una red. Las **Acciones del Defensor** son:

- **Monitorización**: Examina el estado de los nodos para detectar posibles ataques.
- **Aislamiento**: Desconecta un nodo comprometido de la red principal para evitar la propagación del ataque.
- **Redirección a Honeynet**: Engaña al atacante enviando tráfico malicioso a una red ficticia.
- **Actualización de Seguridad**: Reduce la vulnerabilidad de un nodo mediante mejoras en la configuración.
- **Reubicación de Datos Sensibles**: Cambia la ubicación de archivos críticos para evitar su compromiso.

Las **Acciones del Atacante** son:

- **Exploración**: Descubre la topología de la red y las conexiones entre nodos.
- **Escaneo de Vulnerabilidades**: Identifica debilidades en los sistemas objetivos.
- **Explotación de Accesos**: Toma control de un nodo.
- **Escalada de Privilegios**: Intenta obtener acceso administrativo en un nodo controlado.
- **Exfiltración de Datos**: Sustrae información crítica del sistema objetivo.

III-B. Espacio de Observaciones

Cada agente obtiene una representación parcial del estado de la red: **El Defensor** observa:

- Estado de los nodos (atacados o normales).
- Nodos aislados o enviados a la honeynet.
- Ubicación de los archivos sensibles.
- Tipo de servicio de cada nodo.

El Atacante observa:

- Topología descubierta de la red.
- Nodos en los que ha obtenido acceso.
- Estado de los nodos explorados y atacados.
- Éxito o fracaso en intentos de escalada de privilegios y exfiltración de datos.

IV. Metodología y Evaluación

Para evaluar la efectividad de CyberShield, se realizaron 100 simulaciones con distintos tipos de agentes. Se compararon dos estrategias:

- **Estrategia Aleatoria**: Los agentes seleccionan acciones sin una lógica específica.
- **Estrategia Heurística**: Los agentes toman decisiones basadas en reglas predefinidas para optimizar su desempeño.

Los resultados mostraron que los agentes con estrategias heurísticas lograron tasas de éxito significativamente mayores. En particular, los agentes defensores que emplearon estrategias heurísticas lograron un 78 % de victorias, mientras que los agentes atacantes con heurísticas obtuvieron un 62 % de éxito contra defensas aleatorias.

V. Conclusión y Líneas Futuras

Las principales ventajas de CyberShield incluyen:

- Adaptabilidad para nuevos tipos de ataques y defensas.
- Posibilidad de entrenamiento interactivo con humanos.
- Posibilidad de definir distintas configuraciones de red.
- Entorno controlado para evaluar estrategias de IA.

Para mejorar CyberShield, se plantean las siguientes líneas de investigación futura:

- **Transición a un Entorno de Emulación**: Evaluar a los agentes en una infraestructura más realista.
- **Ampliación del Espacio de Acciones**: Incorporar ataques avanzados como DDoS y Man-In-The-Middle, y respuestas defensivas como IDS y firewalls adaptativos.
- **Optimización de Agentes**: Comparar distintas arquitecturas de RL para mejorar la eficiencia de los agentes.

Referencias

[1] J. Á. F. Carrasco, I. A. Pagola, R. O. Urrutia, and R. Román, "Cybershield: A competitive simulation environment for training ai in cybersecurity," in *2024 11th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*. IEEE, 2024, pp. 11–18.

[2] M. D. R. Team., "Cyberbattlesim," https://github.com/microsoft/cyberbattlesim, 2021, created by Christian Seifert, Michael Betser, William Blum, James Bono, Kate Farris, Emily Goren, Justin Grana, Kristian Holsheimer, Brandon Marken, Joshua Neil, Nicole Nichols, Jugal Parikh, Haoran Wei.

[3] A. Andrew, S. Spillard, J. Collyer, and N. Dhir, "Developing optimal causal cyber-defence agents via cyber security simulation," in *Workshop on Machine Learning for Cybersecurity (ML4Cyber)*, 07 2022.

[4] J. Schwartz and H. Kurniawatti, "Nasim: Network attack simulator," https://networkattacksimulator.readthedocs.io/, 2019.

[5] J. Á. Fernández-Carrasco, L. Segurola-Gil, F. Zola, and R. Orduna-Urrutia, "Security and 5g: Attack mitigation using reinforcement learning in sdn networks," in *2022 IEEE Future Networks World Forum (FNWF)*. IEEE, 2022, pp. 622–627.

[6] Y. Huang, L. Huang, and Q. Zhu, "Reinforcement learning for feedback-enabled cyber resilience," *Annual reviews in control*, vol. 53, pp. 273–295, 2022.

[7] Z. HU, "Automated penetration testing using deep reinforcement learning," 2021.

[8] Y. Yang and X. Liu, "Behaviour-diverse automatic penetration testing: A curiosity-driven multi-objective deep reinforcement learning approach," 2022.

[9] M. Standen, M. Lucas, D. Bowman, T. J. Richer, J. Kim, and D. Marriott, "Cyborg: A gym for the development of autonomous cyber agents," *arXiv preprint arXiv:2108.09118*, 2021.

[10] K. Hammar and R. Stadler, "Finding effective security strategies through reinforcement learning and self-play," in *2020 16th International Conference on Network and Service Management (CNSM)*. IEEE, 2020, pp. 1–9.

[11] L. Oakley and A. Oprea, ": An adaptive reinforcement learning strategy for the security game," in *International Conference on Decision and Game Theory for Security*. Springer, 2019, pp. 364–384.

[12] K. Hammar and R. Stadler, "Intrusion prevention through optimal stopping," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 2333–2348, 2022.

Una revisión de: "Desatando la Inteligencia Artificial ofensiva: Generación automática de código de técnicas de ataque"

Eider Iturbe
TECNALIA, Basque Research and Technology Alliance (BRTA)
Derio, Bizkaia
eider.iturbe@tecnalia.com

Oscar Llorente-Vazquez
TECNALIA, Basque Research and Technology Alliance (BRTA)
Derio, Bizkaia
oscar.llorente@tecnalia.com

Angel Rego
TECNALIA, Basque Research and Technology Alliance (BRTA)
Derio, Bizkaia
angel.rego@tecnalia.com

Erkuden Rios
TECNALIA, Basque Research and Technology Alliance (BRTA)
Derio, Bizkaia
erkuden.rios@tecnalia.com

Nerea Toledo
Facultad de Ingeniería, Universidad del País Vasco-Euskal Herriko Unibertsitatea
Bilbao, Bizkaia
nerea.toledo@ehu.eus

Resumen- **La tecnología de Inteligencia Artificial (IA) está revolucionando el mundo digital y se está convirtiendo en la piedra angular de los sistemas digitales modernos. Sin embargo, esta misma tecnología puede ser utilizada como un arma cibernética ofensiva para crear ataques cibernéticos sofisticados y complejos. El trabajo demuestra empíricamente que la IA generativa, concretamente la tecnología de Modelos de Lenguaje Extensos o *Large Language Models* (LLM), puede ser fácilmente utilizada para complementar las plataformas y marcos de Simulación de Brechas y Ataques (BAS) que automatizan la evaluación de seguridad de manera controlada. BAS podría mejorarse con simulaciones de ataques impulsadas por IA, aportando nuevas formas de programar automáticamente múltiples técnicas de ataque, incluso múltiples versiones de la misma técnica de ataque. Por lo tanto, la simulación de ataques mejorada con IA puede ayudar a garantizar que los sistemas digitales sean a prueba de balas y estén protegidos contra una gran variedad de vectores y acciones de ataque.**

***Index Terms*- Inteligencia Artificial, IA Generativa, LLMs, ciberataques, técnicas de ataque.**

Tipo de contribución: *Investigación ya publicada* [1]

I. INTRODUCCIÓN

La Inteligencia Artificial (IA) está transformando el panorama digital y se está integrando cada vez más en la ciberseguridad para automatizar tareas de seguridad y apoyar a los expertos en la toma de decisiones. Sin embargo, las capacidades de la IA también pueden ser empleadas como un arma para crear ciber-ataques avanzados. Este trabajo demuestra empíricamente que la tecnología de Modelos de Lenguaje Extensos (LLM) puede ser utilizada para ayudar en la ejecución automática de ataques.

II. TRABAJOS RELACIONADOS

Con la aparición de soluciones basadas en LLMs están surgiendo nuevos métodos de generación de ataques. Los LLM han demostrado ser efectivos en la generación de código válido, como se describe, por ejemplo, en Chen et al. [2] y Huang et al. [3]. El equipo de Investigación de Check Point utilizó las capacidades de generación de código de ChatGPT y Codex de OpenAI para desarrollar un ataque completo que comienza con un vector de ataque de phishing [4]. Después de numerosos intentos, lograron crear un código exitoso que incluía las acciones necesarias para el ataque. De hecho, el mismo grupo de investigación informó que observaron evidencia de ciberdelincuentes utilizando ChatGPT con fines maliciosos en la Dark Web [5]. Un caso observado involucró a un experto técnico demostrando la creación basada en ChatGPT de varias técnicas de malware comunes.

Algunos trabajos recientes han experimentado con el uso de LLM para la generación de código de ataque. Derner y Batistič [6] estudiaron los riesgos del uso de ChatGPT, incluyendo un ejemplo de generación de código malicioso a través de la falsificación de un prompt de prueba de vulnerabilidad autorizada. En nuestro trabajo, no solo demostramos la viabilidad de la generación de código de identificación de vulnerabilidades, sino también la viabilidad de la automatización completa de múltiples TTP atómicos.

III. METODOLOGÍA

El primer paso de la generación de código de ataques basada en IA consiste en utilizar LLMs para generar el texto del código shell que contiene los comandos para ejecutar la técnica de ataque. El LLM seleccionado es GPT 3.5 debido a su uso ampliamente extendido en el momento de la investigación. El objetivo era solicitar a GPT 3.5 que generara código ejecutable asociado con cada una de las técnicas de ataque definidas por MITRE ATT&CK Enterprise 12.1, que proporcionó la información necesaria para la generación de código.

El prompt se generó empleando características de las técnicas de ataque definidas en el modelo MITRE ATT&CK utilizado y solicitamos generar código shell para esas características específicas. Algunas de las principales características de las técnicas de ataque han influido en la generación del código de la siguiente manera:

- Subtécnica: Es una parte de una técnica más amplia.
- Plataforma: La infraestructura en la que se ha observa-

do la técnica de ataque: Windows, macOS, Linux, Azure AD, Office 365, Google Workspace, etc. Analizando los sistemas operativos más comúnmente utilizados en computadoras de escritorio y servidores, se seleccionaron Windows y Linux en nuestra investigación.

- Táctica: Éstas especifican el propósito detrás de una técnica. Representan el objetivo estratégico del adversario y explican la lógica detrás de sus acciones.

La investigación experimentó con OpenAI playground para determinar el prompt óptimo para pedir a GPT 3.5 que generara código. El código de ataque generado en el paso anterior se ejecutó en ocho máquinas diferentes dentro de un entorno controlado (sandbox). Los sistemas operativos utilizados para el experimento fueron los siguientes: Windows Server 2016, Windows Server 2019, Windows 7, Windows 10, Windows 11, Ubuntu 18.04, Ubuntu 20.04 y Ubuntu 22.04. Durante el proceso, se capturaron los siguientes elementos como evidencia de la ejecución: marca de tiempo de la ejecución, código de retorno, salida estándar, error estándar, mensaje de excepción en caso de que ocurriera, y cualquier binario generado automáticamente por el código.

IV. EVALUACIÓN Y RESULTADOS

Se definieron cinco categorías de logro para caracterizar los resultados de la ejecución del código de ataque generado:

a. *Exitoso*: El código de ataque generado se ajusta a lo que se solicitó originalmente y la ejecución del código en la máquina víctima produce el impacto esperado.
b. *Efectivo - Requiere pequeñas actualizaciones*: El código generado se ajusta a lo solicitado originalmente, pero requiere pequeñas actualizaciones como la configuración de parámetros como nombres de carpetas.
c. *Efectivo dentro de límites - Requiere grandes actualizaciones*: El código generado se ajusta a lo que se solicitó originalmente, pero requiere grandes actualizaciones y/o depende de software de terceros como malware existente.
d. *No funciona*: El código generado se ajusta a lo que se solicitó originalmente pero no funciona correctamente.
e. *Código incorrecto*: El código generado no se ajusta a lo que se solicitó originalmente.

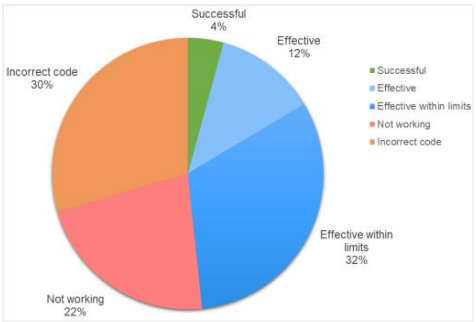


Fig. 1. Resultados de la ejecución del código generado respecto al total del código generado.

Del total de 565 fragmentos de código generados, 362 para Windows y 203 para Linux, los resultados (Fig. 1) mostraron que el 16% de los códigos generados fueron exitosos (sin modificación) o efectivos con pequeñas modificaciones. Y el 32.32% fue efectivo, aunque necesitó grandes actualizaciones. Sin embargo, el 51.68% no funcionaron o fueron incorrectos. Según los resultados por plataforma (Fig. 2), el código generado para Linux fue más exitoso en general en todas las tácticas, excepto en "persistencia" (TA0003) y "escalada de privilegios" (TA00004). La investigación reveló que en nuestro enfoque no hay necesidad de utilizar "jailbreak" para engañar a GPT 3.5 y obligarle a generar código malicioso, dado que en el experimento la estrategia seguida para generar un prompt efectivo fue diferente a dichas técnicas engañosas y GPT 3.5 nunca bloqueó el prompt.

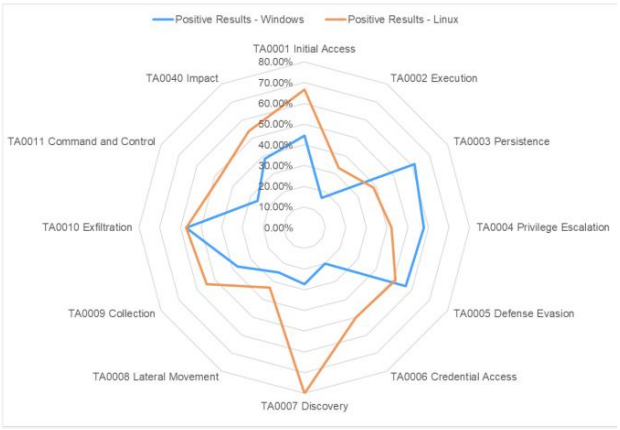


Fig. 2. Comparación de resultados entre plataformas Windows y Linux para las ejecuciones exitosas y efectivas de código generado (categorías a, b, c).

V. CONCLUSIONES

Los resultados validan la hipótesis de que la IA generativa puede mejorar la capacidad de automatizar la ejecución de ciberataques con un esfuerzo mínimo por parte del atacante y sin necesidad de conocimientos previos en programación de técnicas de ataque. Sin embargo, la variabilidad en la generación de código sugiere que se necesita más investigación para optimizar el uso de LLMs en simulación de ciberataques. Los experimentos que se están llevando en la actualidad con LLMs más modernos mejoran los resultados presentados.

AGRADECIMIENTOS

Este trabajo ha recibido financiación del Programa de investigación e innovación Horizonte Europa de la Unión Europea bajo el acuerdo de subvención Núm. 101070450.

REFERENCIAS

[1] Iturbe, E., Llorente-Vazquez, O., Rego, A., Rios, E., & Toledo, N. (2024). Unleashing offensive artificial intelligence: Automated attack technique code generation. Computers & Security, 147, 104077.
[2] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.
[3] Huang, D., Bu, Q., Cui, H., 2023. CodeCoT and beyond: Learning to program and test like a developer. arXiv preprint arXiv:2308.08784.
[4] Ben-Moshe, S., Gekker, G., Cohen, G., 2022. OpwnAI: AI that can save the day or HACK it away. In: Check Point Research. https://research.checkpoint.com/2022/opwnaiai-that-can-save-the-day-or-hack-it-away/. [Accessed: 13-May-2025].
[5] Checkpoint.com. https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/. [Accessed: 13-May-2025].
[6] Derner, E., Batistič, K., 2023. Beyond the safeguards: Exploring the security risks of ChatGPT. arXiv preprint arXiv:2305.08005.

Hacia la integración autónoma de ATT&CK y marcos de seguridad mediante el uso de Agentic AI

Carlos Mario Braga
Grupo de investigación Alarcos
Universidad de Castilla-La Mancha
Ciudad Real, España
CarlosMario.Braga1@alu.uclm.es

Manuel A. Serrano
Grupo de investigación Alarcos
Universidad de Castilla-La Mancha
Ciudad Real, España
Manuel.Serrano@uclm.es

Eduardo Fernández-Medina
Grupo de investigación GSyA
Universidad de Castilla-La Mancha
Ciudad Real, España
Eduardo.FdezMedina@uclm.es

***Resumen*—La correlación eficiente de amenazas y vulnerabilidades sigue siendo un desafío persistente en ciberseguridad, debido a la dispersión de información en múltiples bases de datos, la escasa interoperabilidad entre fuentes y la falta de mecanismos de validación formal. Estas limitaciones impactan directamente en la capacidad de los sistemas para adaptarse a amenazas emergentes, dificultando la contextualización operativa y la trazabilidad de las decisiones.**

Este trabajo presenta una doble contribución. En primer lugar, se lleva a cabo una revisión sistemática de la literatura que clasifica y analiza las principales propuestas en torno a la integración de la base MITRE ATT&CK con otras fuentes de conocimiento clave como CVE, CWE, CAPEC y CPE, considerando además el papel de técnicas generativas como RAG (Retrieval Augmented Generation) e *Agentic AI*. En segundo lugar, se propone una arquitectura basada en *Agentic AI* que automatiza la recuperación, correlación y validación de inteligencia de amenazas en tiempo casi real. Esta arquitectura incorpora un agente orquestador, agentes especializados por fuente, y un mecanismo de verificación basado en *Proof-Carrying Code* (PCC) que permite certificar matemáticamente la consistencia de las correlaciones generadas.

A diferencia de enfoques anteriores, la solución propuesta ofrece una integración unificada, contextualizada y verificable, adaptada a entornos dinámicos. Está diseñada para su aplicación en dominios donde la explicabilidad, la actualización continua y el cumplimiento normativo (por ejemplo, el Reglamento Europeo de IA —AI Act— y el marco de gestión de riesgos de IA del NIST —AI RMF) son requisitos operativos fundamentales. La arquitectura es aplicable a escenarios reales como SOCs, infraestructuras críticas o servicios gestionados, donde se requiere una inteligencia adaptable, trazable y verificable de forma independiente.

***Index Terms*—MITRE ATT&CK, *Agentic AI*, Ciberseguridad, CAPEC, CVE, CWE, CPE.**

Tipo de contribución: *Investigación original*

I. Introducción

La ciberseguridad es esencial para la protección de infraestructuras críticas y activos estratégicos, en un contexto donde las amenazas son cada vez más sofisticadas [1, 2]. La inteligencia de amenazas se apoya en bases de conocimiento como MITRE ATT&CK, CVE, CWE, CAPEC y CPE, que documentan tácticas, vulnerabilidades y productos afectados [3–7]. No obstante, su integración sigue siendo limitada por la fragmentación, la falta de interoperabilidad y la ausencia de mecanismos de validación formal [8, 9].

Estudios recientes han empleado técnicas como *machine learning* o grafos de conocimiento para correlacionar parcialmente estas bases [8, 9], pero con restricciones de escalabilidad y actualización dinámica. Asimismo, el uso de *Retrieval-Augmented Generation* (RAG) ha mejorado la recuperación de información, aunque con limitaciones asociadas a su carácter pasivo, la información desactualizada y la falta de validación estructurada [10–12]. Estas limitaciones son especialmente relevantes en sectores que exigen detección proactiva y cumplimiento continuo, como defensa, salud o infraestructuras críticas, donde la trazabilidad de la inteligencia generada es un requisito operativo.

Este trabajo propone una arquitectura basada en *Agentic AI*, una inteligencia artificial capaz de tomar decisiones y ejecutar acciones por sí misma. La solución incluye un agente orquestador que coordina agentes especializados en la recuperación de información desde fuentes externas (ATT&CK, CVE, etc.) e internas (registros de eventos, *honeypots*), y emplea *Proof-Carrying Code* (PCC) para validar formalmente las correlaciones generadas, reduciendo falsos positivos y garantizando consistencia lógica. Esta arquitectura está diseñada para ofrecer integración continua, personalización contextual y verificación auditable, aspectos clave para su aplicabilidad en entornos normativos como el AI Act o el NIST AI RMF.

La Fig. 1 ilustra la articulación entre las bases mencionadas y su vinculación con datos contextuales internos, permitiendo una inteligencia de amenazas más específica y accionable.

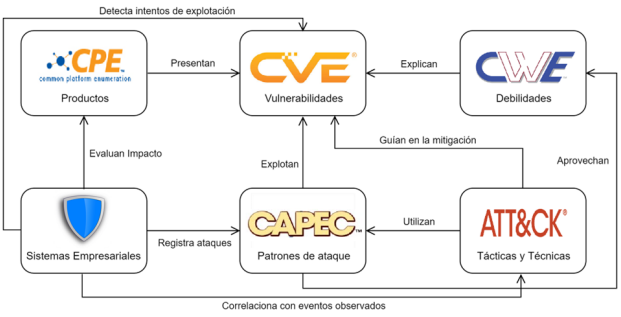


Figura 1. Relación entre bases de conocimiento

El resto del artículo se organiza de la siguiente manera: la Sección II analiza el estado del arte; la Sección III presenta la solución basada en *Agentic AI*; la Sección IV describe la arquitectura y técnicas utilizadas; y la Sección V expone conclusiones y líneas futuras.

Esta propuesta no sólo busca mejorar la eficacia técnica de la inteligencia de amenazas, sino también facilitar su trazabilidad operativa y alineación con marcos de gestión de riesgos de IA, permitiendo avanzar hacia sistemas auditablemente

confiables.

II. Discusión

Para fundamentar esta investigación, se realizó una Revisión Sistemática de la Literatura (SLR) siguiendo el protocolo de Petersen et al. [13]. La pregunta de investigación formulada fue:

> *¿Qué soluciones existen para integrar MITRE ATT&CK con otras bases de ciberseguridad, y hasta qué punto incorporan recuperación automatizada, contextualización operativa y validación de inferencias?*

La búsqueda se llevó a cabo en SCOPUS, IEEE Xplore y ACM Digital Library con la siguiente cadena:

```
('MITRE ATT&CK' OR CAPEC OR CVE OR CWE OR CPE)
AND CYBERSECURITY
AND (RAG OR 'RETRIEVAL-AUGMENTED GENERATION' OR
    'AGENTIC AI')
```

Se recuperaron 95 artículos, de los cuales 35 fueron seleccionados tras aplicar criterios de inclusión y exclusión. Para cada trabajo seleccionado se analizaron el tipo de fuentes integradas, los mecanismos de correlación empleados y la presencia (o ausencia) de validación explícita de resultados. Esta clasificación permitió identificar vacíos comunes y tendencias emergentes en la literatura reciente. Estos trabajos se clasificaron en cuatro líneas principales:

1) Uso de RAG y LLMs (Large Lenguage Models) en ciberseguridad. Varios estudios aplican modelos generativos al análisis de registros de eventos, la detección de incidentes o la verificación de vulnerabilidades [10, 11, 14]. Se propone, por ejemplo, la integración de RAG en SOCs para asistir al personal humano [15], y la combinación con grafos de conocimiento para reducir errores en la interpretación de amenazas [16]. No obstante, estos enfoques dependen de recuperación pasiva, pueden generar alucinaciones y no incorporan validación estructurada [12].

2) Integración parcial de ATT&CK. Se han empleado embeddings [8] y grafos actualizables con LLMs [9] para vincular tácticas con vulnerabilidades como CWE y CVE. Algunos trabajos amplían la correlación con CAPEC y CPE [17, 18], pero siguen centrados en subconjuntos específicos de datos y no integran fuentes empresariales ni garantizan coherencia lógica.

3) Detección de vulnerabilidades con IA. Se han propuesto modelos para desofuscación de malware [19], análisis de código Python [20] y automatización de pentesting [21, 22]. Algunas propuestas comienzan a incluir validación formal mediante técnicas como *Proof-Carrying Code* [23], aunque su adopción es aún incipiente.

4) Defensa autónoma. En entornos industriales y militares se han desarrollado agentes inteligentes capaces de actuar frente a ataques en tiempo real [2, 24]. También se han planteado arquitecturas para simulación de ataques multietapa [25], detección en firmware [1] y análisis de *honeypots* en ICS (Industrial Control Systems) [26]. Sin embargo, ninguna propuesta combina estas capacidades con una integración completa de bases y validación matemática de inferencias.

En conjunto, la literatura revisada refleja avances importantes en tareas parciales —recuperación, detección, simulación—, pero no se ha identificado ninguna propuesta que combine: (i) la integración unificada de ATT&CK, CVE, CWE, CAPEC y CPE; (ii) el uso de datos internos de seguridad como registros de eventos, sensores o *honeypots*; y (iii) un mecanismo de validación formal que garantice coherencia lógica en las inferencias.

Categoría	**Estudios representativos**
RAG y LLMs en ciberseguridad	[10, 11, 14, 16]
Integración de ATT&CK	[8, 9, 12]
Detección de vulnerabilidades con IA	[19–21]
Defensa cibernética autónoma	[1, 2, 26]

Tabla I
Clasificación temática de los estudios revisados.

Además, ninguna solución establece un esquema claro de trazabilidad y verificación alineado con marcos normativos emergentes como el *AI Act* europeo o el *NIST AI RMF*, los cuales exigen explicabilidad técnica, control de riesgos y gobernanza verificable en sistemas de IA sensibles.

Frente a estas carencias, este trabajo propone una arquitectura basada en *Agentic AI*, donde un agente orquestador gestiona la recuperación semántica desde fuentes públicas y empresariales mediante agentes especializados. A esto se añade un componente de validación formal basado en *Proof-Carrying Code*, capaz de certificar que las correlaciones entre tácticas, vulnerabilidades y técnicas de explotación cumplen con invariantes expresadas en lógica de primer orden derivadas de bases como ATT&CK o CAPEC. Este enfoque ofrece un salto cualitativo en precisión, auditabilidad y adaptación dinámica frente a amenazas emergentes.

III. Planteamiento de la solución

La solución propuesta ofrece una integración unificada y dinámica de bases de datos de ciberseguridad y fuentes internas de seguridad empresarial, permitiendo la correlación automatizada de amenazas en tiempo casi real. Su enfoque basado en *Agentic AI* proporciona una serie de beneficios clave frente a los enfoques existentes.

Uno de los aspectos fundamentales es su capacidad de actualización dinámica sin necesidad de reentrenamiento manual. En soluciones previas, cualquier modificación en las bases de datos de ciberseguridad requería ajustes manuales o regeneración de modelos, lo que introducía retrasos en la detección de nuevas amenazas. En contraste, el agente orquestador coordina la recuperación de información de múltiples fuentes en tiempo real, asegurando que la inteligencia de amenazas se mantenga continuamente actualizada y contextualizada.

Otro beneficio fundamental es la precisión en la correlación de amenazas con vulnerabilidades y patrones de ataque. El sistema no solo integra información de MITRE ATT&CK, CVE, CWE, CAPEC y CPE, sino que también incorpora datos internos de la organización, como registros de seguridad, sensores de red y *honeypots* u otras fuentes relevantes, enriqueciendo el contexto operativo de la correlación. Esto permite una detección más alineada con la realidad del entorno, reduciendo falsos positivos y mejorando la capacidad de respuesta ante incidentes.

Además de la mejora en la recuperación y contextualización de amenazas, la solución incorpora un mecanismo de validación formal mediante PCC. Este componente garantiza que las inferencias generadas sean coherentes y verificables, reduciendo la propagación de relaciones incorrectas entre tácticas de ataque y vulnerabilidades. La combinación de *Agentic AI* y PCC permite generar inteligencia de amenazas más fiable y accionable.

Finalmente, el sistema ha sido diseñado para maximizar la escalabilidad y adaptabilidad, evitando dependencias rígidas en reglas predefinidas. A medida que se incorporan nuevas fuentes de datos o cambian los patrones de ataque, el agente orquestador ajusta dinámicamente las consultas y correlaciones sin requerir modificaciones manuales, facilitando la integración de bases de datos adicionales y mejorando la resiliencia del sistema.

La Fig. 2 presenta la arquitectura conceptual de la solución, organizada en varias capas. En la base se encuentran las fuentes de datos, incluyendo bases de ciberseguridad y fuentes internas de la organización. Sobre esta capa, *Agentic AI* gestiona la recuperación, correlación y contextualización de amenazas, mientras que PCC garantiza la validación formal de las inferencias generadas. Finalmente, la capa de presentación y análisis proporciona inteligencia procesada y priorizada a los analistas de seguridad, optimizando la toma de decisiones.

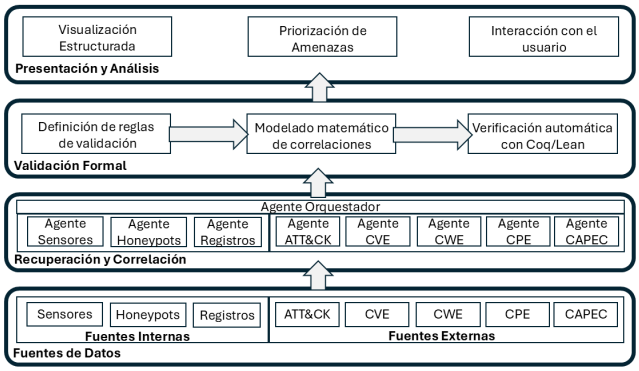


Figura 2. Arquitectura conceptual de la solución propuesta.

Este enfoque proporciona una solución integral para la correlación dinámica y validación de amenazas en ciberseguridad, superando las limitaciones de enfoques tradicionales mediante **Agentic AI**, validación formal y personalización con información contextualizada. En la siguiente sección, se detalla la arquitectura técnica del sistema, describiendo sus componentes y metodologías empleadas para la correlación y validación de amenazas.

IV. Detalles de la solución

La Fig. 2 presenta la arquitectura del sistema propuesto, estructurada en varias capas que interactúan de manera orquestada para recuperar, correlacionar y validar información de ciberseguridad en tiempo casi real.

En la base del sistema se encuentran las fuentes de datos, que incluyen bases de conocimiento ampliamente utilizadas (MITRE ATT&CK, CVE, CWE, CAPEC y CPE), junto con fuentes internas de la organización como sensores de red, *honeypots* y registros de seguridad empresariales. Estas fuentes proporcionan la información primaria que será procesada en las capas superiores. Para cada base de datos pública, se ha definido una estrategia de integración optimizada, asegurando acceso eficiente y actualización continua. La Tabla II resume los métodos de acceso y sincronización utilizados en cada una. Además, los recursos empresariales ofrecen una visión contextualizada del entorno de amenazas, permitiendo correlacionar intentos de ataque detectados en la organización con vulnerabilidades conocidas y tácticas documentadas en ATT&CK.

El agente orquestador, pieza central del sistema, coordina la recuperación de información y la correlación semántica de amenazas en tiempo casi real. Cuando se recibe una consulta, este agente interpreta la solicitud, identifica qué fuentes de datos son relevantes y delega la ejecución de búsquedas a una serie de agentes especializados, cada uno optimizado para interactuar con una base de datos o recurso específico. Además de los agentes que interactúan con bases de datos externas, el sistema incorpora agentes que procesan datos internos de la organización. Por ejemplo, un agente de registros de seguridad puede analizar registros de eventos de tráfico de red para identificar eventos sospechosos, mientras que un agente de *honeypots* correlaciona indicadores de ataque con vulnerabilidades documentadas. Cada agente traduce las consultas del Orquestador en una solicitud técnica adecuada para la API o protocolo de acceso correspondiente y devuelve los resultados estructurados.

Una vez recuperada la información, el agente orquestador fusiona los datos de todas las fuentes y los somete a una validación formal mediante PCC, asegurando la coherencia y precisión de las inferencias generadas. Este mecanismo de validación se estructura en tres fases:

- Definición de reglas de validación: Se establecen invariantes lógicas y propiedades formales que toda correlación debe cumplir. Estas reglas son definidas en lógica de primer orden (FOL) y se derivan de relaciones documentadas en bases de datos como ATT&CK y CAPEC.
- Generación de pruebas formales: Cada correlación generada se modela como una expresión matemática que describe la relación entre una táctica de ataque, una vulnerabilidad y un patrón de explotación. Utilizando teoría de tipos dependientes, el sistema genera una prueba de validez que debe ser resuelta por un verificador automático de pruebas como Coq o Lean.
- Verificación automática: Se ejecuta la prueba formal para garantizar que la correlación generada es consistente con las reglas establecidas. Si la prueba no se valida, la correlación es descartada o marcada para revisión manual. Este proceso reduce falsos positivos y garantiza que las inferencias generadas sean matemáticamente verificables antes de ser presentadas al usuario.

Tras la validación de las correlaciones, la información es organizada en una capa de presentación y análisis, donde se estructura en reportes detallados y se priorizan alertas según su relevancia y criticidad. La presentación de los resultados sigue tres principios clave:

- Visualización estructurada: Se organiza la información

Base de Datos	Formas de Integración Disponibles	Estrategia de Integración
MITRE ATT&CK	Servidor TAXII 2.1, STIX 2.0	TAXII para consultas en tiempo real, STIX para almacenamiento local y procesamiento eficiente
CVE	API RESTful de la NVD, Feeds de datos JSON	API para consultas específicas, Feeds JSON para sincronización local
CWE	Descarga manual de XML/CSV, No tiene API oficial	Descarga periódica y procesamiento local de los archivos XML/CSV
CAPEC	Repositorio CTI de MITRE en STIX 2.0	Uso de la biblioteca stix2 en Python para consultas y almacenamiento
CPE	API RESTful de la NVD, Feeds de datos JSON	API para consultas en tiempo real, Feeds JSON para mantener una base de datos interna actualizada

Tabla II
RESUMEN DE ESTRATEGIAS DE INTEGRACIÓN PARA CADA BASE DE DATOS

en un modelo unificado que permite explorar tácticas, técnicas, vulnerabilidades y patrones de ataque.
- Priorización de amenazas: Se utilizan criterios objetivos, como la criticidad CVSS (Common Vulnerability Scoring System) de una vulnerabilidad, la frecuencia de explotación observada en ATT&CK y el contexto organizacional.
- Interacción con el usuario: Se permite realizar consultas en lenguaje natural, recibir explicaciones detalladas sobre las correlaciones generadas y exportar información a sistemas SIEM (Security Information and Event Management) para su integración en flujos operativos de seguridad.

Este enfoque garantiza que los analistas de seguridad puedan interpretar la información de manera eficiente y tomar decisiones informadas sin necesidad de navegar manualmente por múltiples bases de datos. Con esta arquitectura, el sistema permite una recuperación y correlación de amenazas altamente contextualizada, validada matemáticamente y adaptable a entornos dinámicos, asegurando una inteligencia de amenazas precisa, verificable y escalable.

De forma sucinta y un ejemplo sencillo del funcionamiento del sistema sería cuando un analista de seguridad detecta actividad sospechosa en un servidor y consulta al sistema: '¿Esta actividad está asociada a alguna táctica documentada en MITRE ATT&CK?'. El agente orquestador recupera datos de registros internos y los correlaciona con ATT&CK, CVE y CAPEC, identificando que el evento coincide con la técnica T1071 (Exfiltración a través de Protocolo de Aplicación). Además, se valida la correlación con PCC y se prioriza la alerta según la criticidad del CVE asociado. El sistema recomienda medidas de mitigación y genera un informe exportable para SIEM.

V. CONCLUSIÓN

Este trabajo contribuye con (i) una SLR que clasifica el estado del arte en la integración de MITRE ATT&CK con bases de datos de ciberseguridad, y (ii) una arquitectura *Agentic AI* y validación formal mediante PCC, que tiene como objetivo automatizar la recuperación, correlación y validación de inteligencia de amenazas en tiempo casi real.

Los resultados de la revisión sistemática de la literatura evidencian que ningún trabajo previo ha abordado la integración completa de MITRE ATT&CK, CVE, CWE, CAPEC, y CPE con fuentes de datos empresariales personalizables según la organización, garantizando al mismo tiempo una validación estructurada de las correlaciones generadas. Asimismo, tras revisar soluciones empresariales como Bitdefender y Pentera, tampoco se ha identificado una implementación con estas características. En contraste, la propuesta presentada automatiza la actualización del conocimiento sobre amenazas, reduce la dependencia de reentrenamiento manual y minimiza la propagación de inferencias erróneas, optimizando la inteligencia de amenazas en entornos dinámicos.

A nivel práctico, esta arquitectura ofrece varias ventajas clave. En primer lugar, la automatización y escalabilidad, ya que el uso de *Agentic AI* permite adaptar el sistema a nuevas bases de datos sin modificaciones manuales. En segundo lugar, se logra una mayor precisión en la correlación de amenazas, gracias a la integración de fuentes externas e internas que mejora el análisis contextual de vulnerabilidades. Por último, la validación matemática de inferencias mediante PCC garantiza que las correlaciones generadas sean lógicamente consistentes y verificables.

Dado que este trabajo es una propuesta conceptual, se identifican varias líneas futuras de investigación para avanzar en su desarrollo e implementación. En primer lugar, se propone un piloto del validador formal basado en lógica de primer orden y teoría de tipos dependientes para garantizar la coherencia y precisión de las inferencias, para proseguir con la evaluación empírica del sistema mediante la implementación de un prototipo funcional para analizar su rendimiento en entornos controlados. Posteriormente, se sugiere la optimización del agente orquestador explorando las arquitecturas de agentes más eficientes para mejorar la coordinación entre agentes especializados.

La propuesta presentada representa un avance significativo en la inteligencia de amenazas, proporcionando un modelo escalable, verificable y adaptable para la correlación de datos de ciberseguridad en entornos dinámicos y de alta criticidad. Su implementación futura podría transformar la manera en que las organizaciones gestionan y validan información de amenazas, permitiendo una detección y respuesta más precisa ante actores maliciosos cada vez más sofisticados.

AGRADECIMIENTOS

Este trabajo ha sido desarrollado con el apoyo de los siguientes proyectos: Di4SPDS (PCI2023145980-2), financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea (Programa Chist-Era),

AETHER-UCLM (PID2020-112540RB-C42) financiado por MCIN/AEI/10.13039/501100011033, ALBA (TED2021-130355B-C31) financiado por MICIU/AEI /10.13039/501100011033 y la Unión Europea NextGeneration EU/ PRTR, y MESIAS (2022-GRIN-34202) financiado por FEDER.

Referencias

[1] M. Beninger *et al.*, "ERS0: Enhancing military cybersecurity with AI-driven SBOM for firmware vulnerability detection and asset management," in *2024 16th International Conference on Cyber Conflict: Over the Horizon (CyCon)*. IEEE, 2024, pp. 141–160.

[2] J. F. Loevenich *et al.*, "Training autonomous cyber defense agents: Challenges & opportunities in military networks," in *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*. IEEE, 2024, pp. 158–163.

[3] MITRE, "MITRE ATT&CK: Tactics, techniques and procedures," https://attack.mitre.org/, 2025, accessed on 01 March 2025.

[4] ——, "Common vulnerabilities and exposures (CVE)," https://www.cve.org/, 2025, accessed on 01 March 2025.

[5] ——, "Common weakness enumeration (CWE)," https://cwe.mitre.org/, 2025, accessed on 01 March 2025.

[6] ——, "Common attack pattern enumeration and classification (CAPEC)," https://capec.mitre.org/, 2025, accessed on 01 March 2025.

[7] National Institute of Standards and Technology (NIST), "Common platform enumeration (CPE)," https://nvd.nist.gov/products/cpe, 2025, accessed on 01 March 2025.

[8] Y. Li, L. Zhang, W. Wang, and K. Xu, "AWEB: Bridging cybersecurity attack patterns and weaknesses," in *2023 IEEE International Conference on Communications (ICC)*. IEEE, 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10825621

[9] C. Mishra *et al.*, "PageLLM: Incremental approach for updating a security knowledge graph by using page ranking and large language model," *Information Processing & Management*, vol. 62, no. 3, p. 104045, 2025.

[10] M. Sewak, V. Emani, and A. Naresh, "CRUSH: Cybersecurity research using universal LLMs and semantic hypernetworks," in *Proceedings of the Workshop on Enterprise Knowledge Graphs using Large Language Models (EKG-LLM 2023) co-located with 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023)*. CEUR-WS.org, 2023, pp. 5–15. [Online]. Available: https://ceur-ws.org/Vol-3532/paper5.pdf

[11] Š. Balogh, M. Mlynček, O. Vraňák, and P. Zajac, "Using generative AI models to support cybersecurity analysts," *Electronics*, vol. 13, no. 23, p. 4718, 2024.

[12] Y. Zhang, T. Du, Y. Ma, X. Wang, Y. Xie, G. Yang, Y. Lu, and E.-C. Chang, "AttacKG+: Boosting attack knowledge graph construction with large language models," *arXiv preprint arXiv:2405.04753*, 2024.

[13] K. Petersen *et al.*, "Systematic mapping studies in software engineering," in *12th international conference on evaluation and assessment in software engineering (EASE)*. BCS Learning & Development, 2008.

[14] T. ZeMicheal, H. Chen, S. Davis, R. Allen, M. Demoret, and A. Song, "LLM agents for vulnerability identification and verification of CVEs," *CEUR Workshop Proceedings*, 2024.

[15] R. Kurnia, F. Widyatama, I. M. Wibawa, Z. A. Brata, G. A. Nelistiani, H. Kim *et al.*, "Enhancing security operations center: Wazuh security event response with retrieval-augmented-generation-driven copilot," *Sensors (Basel, Switzerland)*, vol. 25, no. 3, p. 870, 2025.

[16] L. Ou, X. Ni, W. Wu, and Z. Tian, "CyGPT: Knowledge graph-based enhancement techniques for large language models in cybersecurity," in *2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2024, pp. 216–223.

[17] T. Van Ede, T. Mulder, and E. Moyakine, "Interdisciplinary research project 'AI Shield'," in *2024 8th Cyber Security in Networking Conference (CSNet)*. IEEE, 2024, pp. 268–271.

[18] Y. Li, Q. Sun, K. Wang, and C. Zhou, "Integrated digital model-based safety and security awareness and analysis for process industry," in *2024 China Automation Congress (CAC)*. IEEE, 2024, pp. 1483–1488.

[19] C. Patsakis, F. Casino, and N. Lykousas, "Assessing LLMs in malicious code deobfuscation of real-world malware campaigns," *Expert Systems with Applications*, vol. 256, p. 124912, 2024.

[20] A. Mechri *et al.*, "Secureqwen: Leveraging LLMs for vulnerability detection in python codebases," *Computers and Security*, vol. 148, p. 104151, 2025.

[21] J. Huang and Q. Zhu, "Penheal: A two-stage LLM framework for automated pentesting and optimal remediation," in *Proceedings of the Workshop on Autonomous Cybersecurity*, 2023, pp. 11–22.

[22] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass, "PentestGPT: Evaluating and harnessing large language models for automated penetration testing," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 847–864.

[23] P. Kamran, P. Devanbu, and C. Stanford, "Vision paper: Proof-carrying code completions," in *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering Workshops*, 2024, pp. 35–42.

[24] J. F. Loevenich, E. Adler, R. Mercier, A. Velazquez, and R. R. F. Lopes, "Design of an autonomous cyber defence agent using Hybrid AI models," in *2024 International Conference on Military Communication and Information Systems (ICMCIS)*. IEEE, 2024, pp. 1–10.

[25] Ö. Sen, C. Pohl *et al.*, "AI-Based attacker models for enhancing multi-stage cyberattack simulations in smart grids using co-simulation environments," in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2024, pp. 68–75.

[26] A. H. Nursidiq and C. Lim, "Threat analysis on industrial control system based on attacker's behaviors using honeypots," in *2022 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES)*, 2022, pp. 1–6.

Sesión Pósters II: Privacidad, datos sensibles y tecnologías cuánticas

A review of "Privacy-aware Berrut Approximated Coded Computing applied to Federated Learning"

Xavier Martínez-Luaña*†, Manuel Fernández-Veiga†, Rebeca P. Díaz-Redondo†
*Galician R&D Center in Advanced Telecommunications (GRADIANT), Vigo, Spain
Email: xamartinez@alumnado.uvigo.es, xmartinez@gradiant.org
†atlanTTic Research Center, Information & Computing Lab, Universidade de Vigo, Spain
Email: {mveiga, rebeca}@det.uvigo.es

***Abstract*—Coded computing is currently being used for solving significant challenges in distributed learning systems. These challenges include particularly data privacy during distributed training, where coding is used to conceal information on the model and the training samples to active or passive attackers. However, coded computing is generally designed for exact recovery of the data, and works only for a limited class of functions and for computations on finite alphabets. This paper validates the Private Berrut Approximated Coded Computing as a solution to privacy concerns tailored for general FL settings. PBACC leverages Berrut Approximated Coded Computing to implement a Secret Sharing protocol capable of computing non-linear functions while maintaining a balance between privacy, precision, and complexity. We integrate PBACC into a FL framework for image processing using two popular datasets (MNIST and CIFAR-10), two different aggregation functions (FedAvg and FedMedian), and we deal with two restrictions: honest but curious nodes and malicious nodes. We have performed different tests under these conditions to measure privacy, precision, and complexity to determine the viability of the trade-offs for real-world FL scenarios, concluding this approach offers promising results.**

I. Introduction

Privacy concerns in machine learning have led to the emergence of Federated Learning (FL), a decentralized paradigm where data remains on local devices while models are trained collaboratively. Despite this, FL is vulnerable to inference and reconstruction attacks, specially if the attacker has access to the local model updates of the data owners.

Coded computing has emerged as a potential solution, using coding theory to conceal data and intermediate computations. However, most existing coded computing schemes are designed for exact recovery and operate over finite fields, making them unsuitable for real FL, where real numbers and non-linear functions are commonly used. [1] [2]

In our work, we introduce PBACC (Privacy-aware Berrut Approximated Coded Computing), a coding framework that enables the approximation of non-linear computations while maintaining privacy guarantees. We validate PBACC in a realistic FL scenario for two two different threat models: (i) a minority of colluding nodes that want to infer information about the inputs, and (ii) an extension of the previous one adding malicious nodes that introduce alterations in the local trained models.

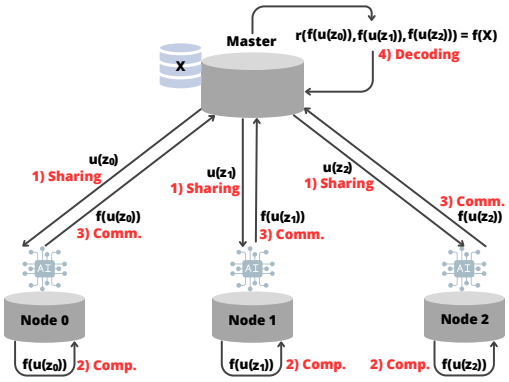


Fig. 1. Decentralized computation of a function using PBACC for 3 nodes.

II. PBACC: Privacy-aware Berrut Approximated Coded Computing

PBACC builds on top of the Berrut Approximate Coded Computing scheme by extending the encoding function to add some random coefficients that will provide privacy to the encoded input, without having a great impact on precision. This enables the evaluation of arbitrary non-linear functions required for general FL.

Suppose a collection of N nodes, each one having a private data vector $\mathbf{X}^{(i)} = \begin{pmatrix} X_0^{(i)} & \dots & X_{K-1}^{(i)} \end{pmatrix}^T$, where the input data $\mathbf{X}^{(i)} \in \mathbb{R}^{K\times L}$ is owned by node i, for $i \in \{0,\dots,N-1\}$. The PBACC computation scheme consists of three phases, that are detailed as follows:

1) Encoding and sharing: The node i constructs the following rational function

$$u_{\mathbf{X}^{(i)}}(z) = \sum_{j=0}^{K-1} \frac{\frac{(-1)^j}{(z-\alpha_j)}}{\sum_{k=0}^{K+T-1}\frac{(-1)^k}{(z-\alpha_k)}} X_j^{(i)} + \sum_{j=0}^{T-1} \frac{\frac{(-1)^{j+K}}{(z-\alpha_j)}}{\sum_{k=0}^{K+T-1}\frac{(-1)^j}{(z-\alpha_k)}} R_j^{(i)}, \quad (1)$$

[1]This work was supported by "TRUstworthy Multi-site Privacy Enhancing Technologies (TRUMPET)" (10.3030/101070038), by the grant PID2020-113795RB-C33 funded by MICIU/AEI/10.13039/501100011033 (COMPROMISE project), "Contramedidas Inteligentes de Ciberseguridad para la Red del Futuro (CICERO)" (CER-20231019) funded by CERVERA grants 2023 for RTOs, by the grant PID2023-148716OB-C31 funded by MCIU/AEI/ 10.13039/501100011033661 (DISCOVERY project), by "TRUFFLES: TRUsted Framework for Federated LEarning Systems, within the strategic cybersecurity projects (INCIBE, Spain), funded by the Recovery, Transformation and Resilience Plan (European Union, Next Generation)", and from the Galician Regional Government under project ED431B668.

[2]X. Martínez-Luaña, M. Fernández-Veiga and R. P. Díaz-Redondo, *"Privacy-aware Berrut Approximated Coded Computing applied to Federated Learning"*, 2024 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1-6, doi: 10.1109/WIFS61860.2024.10810717.

for some distinct regression points $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{K-1}) \in \mathbb{R}^K$ associated to the data $\mathbf{X}^{(i)}$, which we choose as the Chebyshev points of first kind $\alpha_j = \cos\left(\frac{(2j+1)\pi}{2K}\right)$, and some distinct interpolation points $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{T-1}) \in \mathbb{R}^T$ associated to the randomness $R_j^{(i)}$, which are selected as the shifted Chebyshev points of first kind $\alpha_{K+j} = b + \cos\left(\frac{(2j+1)\pi}{2K}\right)$. The rational function $u_{\mathbf{X}^{(i)}}(z)$ is then evaluated at the subset $\{z_j\}$, for $j = 0, \ldots, N-1$, specifically using $z_j = \cos(\frac{j\pi}{N-1})$. Finally, the random coefficients $R_j^{(i)}$ are distinct $1 \times L$ vectors obtained from a Gaussian distribution $\mathcal{N}(0, \frac{\sigma_n^2}{T})$. The evaluation $u_{\mathbf{X}^{(i)}}(z_j)$ is the share created from the node i to be sent to node j.

2) Computation and communication: Node i calculates an arbitrary function f using a set of rational function evaluations shared from the rest of the nodes $\{u_{\mathbf{X}^{(0)}}(z_i), u_{\mathbf{X}^{(1)}}(z_i), \ldots, u_{\mathbf{X}^{(N-1)}}(z_i)\}$, where $u_{\mathbf{X}^{(j)}}(z_i)$ is the evaluation of the polynomial quotient associated to the input $X^{(j)}$ owned by the j-th node, and shared with the node i. The node i computes $f\big(u_{\mathbf{X}^{(j)}}(z_i)\big), \quad j = \{0, \ldots, N-1\}$ and sends the result to the aggregator.

3) Decoding and reconstruction: In the last stage, the aggregator node reconstructs the value of the objective function $f(\cdot)$ over all the inputs using the results obtained from the subset $\mathcal{F}$ of the fastest nodes, computing the Berrut Rational Interpolation:

$$r_{\text{Berrut},\mathcal{F}}(z) = \sum_{i=0}^{n} \frac{\frac{(-1)^i}{(z-\tilde{z}_i)}}{\sum_{j=0}^{n} \frac{(-1)^j}{(z-\tilde{z}_j)}} f\big(v(\tilde{z}_i)\big), \tag{2}$$

where $\tilde{z}_i \in \mathcal{S} = \{\cos\frac{j\pi}{N-1}, j \in \mathcal{F}\}$ are the interpolation points. With this values, the aggregator evaluates the approximation $f(X_j^{(i)}) \approx r_{\text{Berrut},\mathcal{F}}(\alpha_j)$, for all $j \in \{0, \ldots, K-1\}$ and $i \in \{0, \ldots, N-1\}$.

III. Experimental Results

In order to assess the behaviour of our proposal, we have performed different experiments under the following conditions. First, including C honest but curious nodes and up to C malicious nodes. Second, using two different aggregation functions: FedAvg and FedMedian. We implemented a convolutional neural network (CNN) model using two common datasets, MNIST and CIFAR-10. It must be noted that the experiments have been conducted in an emulated setting where all the nodes run on the same computing machine and share memory to exchange the information.

A. CNN on MNIST

As it can be observed in Table I, the round of convergence is quite similar in all configurations. This implies that PBACC is able to provide privacy guarantees while having a very low effect on the normal behaviour of the CNN-MNIST in FedAvg and FedMedian settings.

Regarding times, we observe that the PBACC additional computations add only a small overhead of 5.75 seconds for FedAvg + PBACC and 9.87 for FedMedian + PBACC, per round of FL. Since computations are done sequentially, we can easily calculate that the encoding, aggregations, and decoding are adding approximately 0.11 and 0.19 seconds respectively, to the training time per round of each node. In this case, scenarios have reached a 100% of accuracy. This demonstrates that even if PBACC introduces approximate computation of functions, the privacy/precision trade-off still allows the model to converge.

TABLE I
Comparison of the tested scenarios using MNIST dataset

Convergence Round	FedAvg	FedAvg + PBACC	FedMedian	FedMedian + PBACC
μ	2.94	2.76	2.72	3.06
σ	0.94	0.86	0.77	0.92
Total Time (s)	**FedAvg**	**FedAvg + PBACC**	**FedMedian**	**FedMedian + PBACC**
Per round	24.15	29.90	22.48	32.35
Per FL task	71.02	82.52	61.15	99.00

TABLE II
Comparison of the tested scenarios using CIFAR-10 dataset

Convergence Round	FedAvg	FedAvg + PBACC	FedMedian	FedMedian + PBACC
μ	5.20	8.82	6.2	9.74
σ	2.18	4.29	2.68	5.52
Total Time (s)	**FedAvg**	**FedAvg + PBACC**	**FedMedian**	**FedMedian + PBACC**
Per round	173.01	464.94	202.44	1018.65
Per FL task	899.67	4,100.84	1,255.13	9,921.68

B. CNN on CIFAR-10

As Table II summarizes, the round of convergence for this CNN, in terms of mean and standard deviation, shows worse results than in the MNIST dataset, as the CNN in this case has many more parameters, and it is more sensitive to the error. This implies that PBACC is still able to provide privacy guarantees, but in this case, it will require a few more rounds to converge for both cases.

Regarding times, we can observe that due to the size of this CNN, the PBACC additional computations adds an overhead of 291.93 seconds for FedAvg + PBACC and 816.21 for FedMedian + PBACC, per round of FL. Since computations are done sequentially, we can easily calculate that the encoding, aggregations, and decoding are adding approximately 5.83 and 16.32 seconds respectively, to the training time per round of each node. For the CNN on CIFAR-10, the achieved accuracies are 87% for FedAvg, 86% for FedAvg + PBACC, 87% for FedMedian, and 87% for FedMedian + PBACC. These results are quite interesting, since they show that even if the error introduced by PBACC make the models converge slower, the final accuracy achieved are quite similar, which demonstrates that the PBACC privacy/precision exchange allows to provide privacy guarantees while not affecting the final quality of the model.

IV. Conclusion and Future Work

Given the proimising results obtained, PBACC seems suitable for real-world FL. Future work will address the implementation of well-known attacks in FL environments, and the integration of PBACC into the whole distributed training task, which will increase computational complexity in exchange for reducing communication costs.

A review of "Privacy-preserving protocol for high-frequency smart meters using reversible watermarking and paillier encryption"

Farzana Kabir[1,2], Tanya Koohpayeh Araghi[1,2], and David Megías[1,2]
[1] Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Spain
[2] CYBERCAT-Center for Cybersecurity Research of Catalonia
{fkabir, tkoohpayeharaghi, dmegias}@uoc.edu

***Resumen*—This research presents a summary of the privacy-preserving protocol for high-frequency smart meters (P3HF) using difference expansion-based reversible watermarking and Paillier homomorphic encryption. The proposed protocol improves security and user data privacy while consuming less energy and time. Experiments on an Arduino UNO Rev.3 hardware platform show that the protocol performs better than previous researches in terms of data requirements, privacy resilience, and capacity to overcome limitations.**

***Index Terms*—Smart Meter, Watermarking, Encryption, Security, Privacy.**

Tipo de contribución: Investigación ya publicada [1]

I. Introduction

Smart meters generate a large amount of data, enabling service providers to track and manage energy consumption instantly [2]. Unfortunately, this real-time data transfer poses risks to consumers' security, safety, and privacy. Tampering data in smart meters can cause financial or personal harm [3] and can be used for any type of criminal intent which can be a great threat to user privacy and data integrity [4], [5].

The study presents a secure, privacy-preserving protocol for high-frequency smart meters (P3HF) using reversible digital watermarking and Paillier homomorphic encryption to mitigate the risk of confidential information leakage. The main contributions are as follows.

- A privacy-preserving protocol is proposed for high-frequency smart meters (P3HF) using reversible watermarking and Paillier encryption.
- A unique encryption server is introduced to perform the homomorphic encryption part and reduce the computational burden of smart meters.
- The efficiency and reliability of our proposed protocol in terms of minimal energy and time consumption is proved by implementation of the protocol on Arduino UNO Rev.3.

II. Methodology

The proposed P3HF protocol uses Paillier encryption and reversible watermarking in the four components of the system: Smart Meter (SM), Encryption Server (ES), Data Aggregator (DA), and Control Center (CC) and consists of 5 phases as demonstrated in Figure 1:

When considering a total of n SMs and m time frames, each smart meter SM_i sends its energy consumption data, d_{ij}, at each time period, t_j, where $i = 1, 2, 3, ..., n.$ and $j = 1, 2, 3, ...m.$ SM_i receives the 256-bit key K_w generated by CC for the cryptographic hash function SHA-2 to generate the watermark (W). The watermark W is generated based on SHA-2, $H_j = \text{Hash}(K_w, t_{2j})$ for $j = 1, 2, \ldots, \frac{m}{2}$. To construct the watermark, the individual m hash values are combined using an XOR operation: $W = H_1 \oplus H_2 \oplus \cdots \oplus H_{\frac{m}{2}}$.

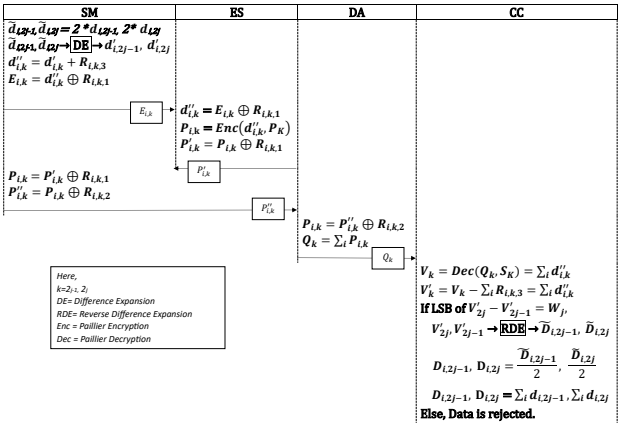


Figura 1. P3HF protocol phases

Each of the n SMs generates three pseudorandom numbers $R_{i,2j,1}$, $R_{i,2j,2}$, and $R_{i,2j,3}$ respectively using and inital $seed$. CC also generates a pair of public and private keys (P_K, S_K) for Paillier encryption and decryption and shares P_K with ES.

To execute the protocol phases, we first divide the m time frames into $\frac{m}{2}$ pairs, which indicates that each iteration considers just 2 time frames. Hence, at each time frame, for all SM_i, the relevant data are $d_{i,2j-1}$ and $d_{i,2j}$, for $j = 1, \ldots, \frac{m}{2}$. The 5 phases of our P3HF protocol are explained below:

- At each even time frame $2j$, one bit of the watermark (W_{2j}) is embedded into the LSB of $(\widetilde{d}_{i,2j-1})$-$(\widetilde{d}_{i,2j})$ to obtain the watermarked data $d'_{i,2j-1}$ and $d'_{i,2j}$. We denote $k = 2j-1, 2j$ in this protocol. The pseudorandom numbers $R_{i,k,3}$ are added to the watermarked data to obtain $d''_{i,k}$. Then we again XOR pseudorandom numbers $R_{i,k,1}$ and get $E_{i,k}$.
- Upon receiving $E_{i,k}$, ES performs XOR to obtain $d''_{i,k}$ and then generates the ciphertext $P_{i,k}$ by applying Paillier encryption using the public key (P_K). To secure

the communication, $R_{i,k}$ is used again to XOR the ciphertext and is transmitted back to SM_1 as $P'_{i,k}$ (Here, $k = 2j-1 and 2j$ so when we write $P'_{i,k}$, we mean there are two data $P'_{i,2j-1}$ and $P'_{i,2j}$).

- SM_i performs an XOR operation using $R'_{i,k,1}$ to obtain $P_{i,k}$. Finally, it again encrypts $P_{i,k}$ using $R_{i,k,2}$ before sending it to DA as $P''_{i,k}$.
- DA receives the watermarked encrypted data $P''_{i,k}$ from all n SM_i and performs the decryption using $R_{i,k,2}$ to obtain $P_{i,k}$. Then, DA aggregates all $P_{i,k}$ together using a simple addition. Finally, these aggregated data, $Q_k = \sum_i P_{i,k}$ are transmitted to CC.
- CC uses its private key (S_K) to decrypt Q_k and obtain $V_k = \sum_i d''_{i,k}$. Then CC subtracts $\sum_i R_{i,k,3}$ from V_k to obtain $V'_k = \sum_i d'_{i,k}$. The watermark is then validated by comparing the LSB of $(V'_{2j} - V'_{2j-1})$ and W_j. This is true only if n is odd. If n is even, the LSB of $V'_{2j} - V'_{2j-1}$ will always be 0. If they are not equal, it is assumed that the data was tampered with and is declined immediately. If the LSB is correct, CC extracts the watermark and computes the sum of the plain text data: $\widetilde{D}_{2j-1} = \sum_i \widetilde{d}_{i,2j-1}$ and $\widetilde{D}_{2j} = \sum_i \widetilde{d}_{i,2j}$. The final aggregated data is obtained simply by dividing by two: $D_{2j-1} = \frac{\widetilde{D}_{2j-1}}{2}, D_{2j} = \frac{\widetilde{D}_{2j}}{2}$.

III. Results

The operations of smart meters are executed using an Arduino UNO Rev3, which is a microcontroller board based on the ATmega328P. We ran the program 100 times, and it averaged 24.14 ms per iteration on Arduino Uno Rev3, which is significantly lower than many of the recent works. To show the effectiveness of P3HF in terms of data expansion, communication cost in bits, and computation cost, we have compared it with four other recent protocols, as indicated in Table I. Figure 2 shows a graph comparing the total communication costs and computation costs at SM for different methods as the number of SMs increases from 1 to 100. P3HF continually performs better than the other approaches in terms of computational cost per SM and total communication costs, demonstrating its superior scalability and efficiency while managing a large number of high-frequency smart meters. As a result, P3HF is the most resource-efficient approach among them.

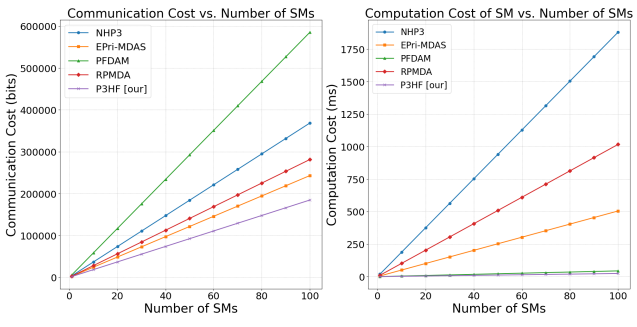


Figura 2. Performance comparison Graph

Tabla I
Performance comparison (n = number of SMs)

Methods	Computation Cost (ms)			Communication Cost (bits)
	SM	DA	CC	
NHP3 [6]	18.8	18.8n + 18.8	18.8	3688
EPri-MDAS [7]	5.05	0.59 + 0.75n	1.86	2432
PFDAM [8]	0.44	18.29n	18.29	5856
RPMDA [9]	10.17	10.17n + 20.34	20.34	2816
P3HF [1] [prop.]	0.24	0.01n	1.54	1848

IV. Conclusion

The study presents P3HF, a privacy-preserving data aggregation protocol for high-frequency smart metering systems. It employs reversible watermarking based on difference expansion and homomorphic Paillier encryption to prevent potential attacks. The protocol is lightweight, efficient, and secure, guaranteeing data integrity and secrecy. It is effective against both active and passive attacks, with minimal communication and computational cost.

Acknowledgments

The authors acknowledge the funding obtained from the "SECURING" project (PID2021-125962OB-C31) funded by the Ministry of Science and Innovation, the Agencia Estatal de Investigación, and the European Regional Development Fund (ERDF), and by the ARTEMISA International Chair of Cybersecurity (C057/23) and the DANGER Strategic Project of Cybersecurity (C062/23), both funded by the Spanish National Institute of Cybersecurity through the European Union – Next GenerationEU and the Recovery, Transformation, and Resilience Plan.

Referencias

[1] F. Kabir, T. K. Araghi, and D. Megías, "Privacy-preserving protocol for high-frequency smart meters using reversible watermarking and paillier encryption," *Computers and Electrical Engineering*, vol. 119, p. 109497, 2024.

[2] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5820–5830, 2017.

[3] M. Nateghizad, Z. Erkin, and R. L. Lagendijk, "An efficient privacy-preserving comparison protocol in smart metering systems," *EURASIP Journal on Information Security*, vol. 2016, no. 1, pp. 1–8, 2016.

[4] Z. Wang, Y. Liu, Z. Ma, X. Liu, and J. Ma, "Lipsg: lightweight privacy-preserving q-learning-based energy management for the iot-enabled smart grid," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3935–3947, 2020.

[5] T. K. Araghi, D. Megías, and A. Rosales, "Toward the selection of a lightweight authentication technique for the security of smart homes: Framework architecture based on a user centric design," in *Science and Information Conference*. Springer, 2023, pp. 651–667.

[6] A. Mohammadali and M. S. Haghighi, "A privacy-preserving homomorphic scheme with multiple dimensions and fault tolerance for metering data aggregation in smart grid," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5212–5220, 2021.

[7] J. Zhang, W. Zhang, X. Wei, and H. Liu, "Epri-mdas: An efficient privacy-preserving multiple data aggregation scheme without trusted authority for fog-based smart grid," *High-Confidence Computing*, p. 100226, 2024.

[8] J. Zhang and J. Wei, "Pfdam: Privacy-preserving fine-grained data aggregation scheme supporting multi-functionality in smart grid," *IEEE Internet of Things Journal*, 2024.

[9] J. Liu, H. Wang, J. Bao, R. Sun, X. Du, and M. Guizani, "Rpmda: Robust and privacy-enhanced multidimensional data aggregation scheme for fog-assisted smart grids," *IEEE Internet of Things Journal*, 2024.

A review of: Enhancing Privacy in Federated Learning: A Practical Assessment of Combined PETs in a Cross-Silo Setting

Jaime Loureiro-Acuña
Xavier Martínez-Luaña
Gradiant
Vigo, Spain
{jloureiro,xmartinez}@gradiant.org

Héctor Padín-Torrente
Gonzalo Jiménez-Balsa
Gradiant
Vigo, Spain
{hpadin,gjimenez}@gradiant.org

Carlos García-Pagán
Ines Ortega-Fernandez
Gradiant
Vigo, Spain
{cgarcia,iortega}@gradiant.org

***Abstract*—Federated learning is a decentralized, collaborative learning paradigm that aims to preserve data privacy during the training process. Recent research, however, has revealed significant privacy attacks that pose a considerable threat to the adoption of Federated Learning as a technology intended to ensure privacy during the training of sensitive data. To tackle these challenges, Privacy-Enhancing Technologies have emerged as effective solutions for protecting data privacy in these environments. This study evaluates and benchmarks two distinct PET approaches for cross-silo Federated Learning: 1) Secure Multi-Party Computation with Homomorphic Encryption, and 2) Coded Distributed Computing with Differential Privacy. Through a comprehensive evaluation, this paper compares the computational efficiency, communication costs, and model accuracy of these techniques. The results highlight the potential of these technologies to effectively mitigate privacy risks, with each method presenting distinct trade-offs between accuracy and performance.**

***Index Terms*—Federated Learning, Privacy Enhancing Technologies, Distributed Computing**

Type of contribution: *Already published [1]*

I. Introduction

Federated Learning (FL) has emerged as a paradigm shift in machine learning, enabling collaborative model training across multiple entities while maintaining data privacy. By keeping data decentralized and sharing only model updates, FL addresses critical privacy concerns inherent to centralized data processing.

Despite its potential, FL implementations are not without vulnerabilities. Recent studies have demonstrated that sensitive information can still be inferred through model updates, and adversarial entities can exploit weaknesses in the learning process. These challenges highlight the urgent need for robust privacy-preserving mechanisms to ensure secure and reliable deployment of FL systems.

Privacy Enhancing Technologies (PETs) offer a promising approach to address these challenges. Despite their potential, the adoption of PETs in practical FL scenarios remains constrained by several challenges. These include data degradation, numerical instability, difficulties in performing specific computations, and substantial overheads. These obstacles highlight the need for continued research and optimization to make PETs more practical and scalable in real-world FL applications.

This paper evaluates the integration of two PETs in a cross-silo FL scenario (Figure 1): (1) Secure Multi-Party Computation (SMPC) with Homomorphic Encryption (HE) and (2) Coded Distributed Computing (CDC) with Differential Privacy (DP). These methods are assessed in terms of training efficiency, and the impact on the final model performance. The results provide valuable insights into the strengths and limitations of each approach, offering guidance for the development of privacy-preserving FL systems.

II. Practical application scenario

To assess the practical applicability of PETs, this study adopts a cross-silo FL scenario. In this setting, Data Owners (DOs), representing distinct organizations, collaboratively train a global logistic regression (LR) model using the "Breast Cancer Wisconsin" oncological dataset. The dataset, comprising 687 entries with 11 features, was partitioned such that each of the five DOs held 110 entries, with the remaining data reserved for model performance evaluation. The LR was trained utilizing Federated Averaging for binary breast cancer classification, across 500 local epochs and 1000 rounds with varied initialization seeds. Our approach was evaluated on a server environment with VM-simulated entities, leveraging a PySyft-based FL platform specifically customized with a communication module to enhance efficiency through parallel message handling. To simulate a realistic threat environment, our model incorporates a dishonest majority assumption, specifically considering up to three colluding DOs exhibiting

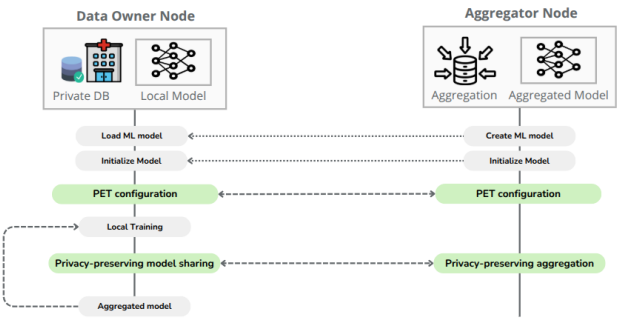


Figure 1. High-Level overview of PET integration for a cross-silo FL scenario

a curious-but-passive behavior.

III. Evaluated PET approaches

A. SMPC and HE

In FL, both SMPC and HE offer pathways to privacy. However, each has inherent limitations when applied independently. Multiparty Homomorphic Encryption (MHE) emerges as a strategic combination of both techniques to mitigate these limitations. MHE leverages the core strength of HE, the ability to compute on encrypted data, to ensure privacy during the aggregation phase in FL. Simultaneously, MHE incorporates elements inspired by SMPC, particularly for distributed protocols. Our implemented MHE scheme comprises the following steps in a cross-silo FL scenario:

1) **Setup** and **Key Generation**: DOs engage in a secure, distributed key generation protocol to establish cryptographic keys without revealing individual secret information.
2) **Local Model Update Encryption**: Each DO encrypts their locally trained model updates using the established encryption key.
3) **Homomorphic Aggregation**: The Aggregator receives the encrypted model updates and performs the aggregation function in the homomorphic domain.
4) **Distributed Decryption**: The DOs engage in a distributed decryption process to collectively decrypt the aggregated model without revealing their individual updates.

By combining HE and SMPC-inspired protocols, MHE aims to achieve a more practical and robust privacy solution for FL, balancing computational efficiency with strong privacy guarantees.

B. CDC and DP

Combining CDC with DP strategically aims to leverage the strengths of both. DP, applied to the local model, ensures that the final global model, and any inferences drawn from it, are differentially private. CDC enables participants to efficiently contribute to the aggregation process by sharing encoded model updates. Our implemented scheme comprises the following steps in a cross-silo FL scenario:

1) **Noise Addition**: DOs apply the Sparse Vector Technique with Selective Parameter Update to obtain noisy model updates.
2) **Encoding** : The noisy model updates are encoded using a Lagrange Coded Computing encoding function.
3) **Sharing**: DOs engage in a sharing step, ensuring that each participant holds a necessary set of shares, which is essential for the subsequent evaluation.
4) **Evaluation**: Each DO aggregates the received shares to compute a partial aggregation.
5) **Result Sharing**: DOs send the partial encoded aggregations to the Aggregator.
6) **Decoding**: Outputs the decoded and noisy aggregated model, using polynomial interpolation.

The combination of both techniques provides a practical alternative for privacy-preserving FL, particularly when computational efficiency and control over output privacy are key considerations.

Table I
Comparative of the quality of inference and training times for the three schemes: FL without PETs, FL with MHE and FL with CDC+DP

Scheme		Acc. (%)	Prec. (%)	Recall (%)	AUC-ROC Score	Training Time (ms)
FL	μ	93.62	92.38	89.96	0.969	400.4
	σ	1.1	1.65	1.72	0.009	64.79
MHE	μ	93.62	92.38	89.95	0.969	5895.38
	σ	1.11	1.66	1.72	0.009	383.77
CDC+ DP	μ	88.28	82.35	87.19	0.914	614.36
	σ	3.38	6.07	6.78	0.044	68.28

IV. Benchmarking Results

A comprehensive evaluation of the two proposed PET combinations in our FL platform, is benchmarked in Table I. This table provides a detailed summary of inference quality metrics—Accuracy, Precision, Recall, and AUC-ROC—and the total training time for each scheme under consideration. These findings reveal that the MHE scheme demonstrates inference quality comparable to the baseline FL approach, which operates without PETs. This similarity in performance indicates minimal utility degradation despite the incorporation of privacy mechanisms. It is also evident, however, that MHE introduces a significant increase in total training time. Conversely, the CDC+DP scheme exhibits a somewhat reduced inference quality in comparison to both FL and MHE. Despite this slight reduction in quality, CDC+DP offers a considerably lower training time overhead than MHE. These findings underscore the inherent trade-offs between privacy guarantees, model utility, and computational cost that are intrinsic to different PET approaches in FL.

V. Conclusions

The presented evaluation highlights the promise of two distinct PET combinations in addressing privacy challenges in cross-silo FL. Our evaluation demonstrates that both methodologies offer feasible pathways for privacy-preserving collaborative learning in such environments. While MHE exhibits a higher inference quality, its computational demands for global model generation are considerably higher in comparison to CDC+DP and plain FL. Notably, CDC+DP achieves output privacy through noise injection, ensuring data privacy in the final model. However, the MHE approach ensures that the Aggregator remains uninformed about the final model in its decrypted state.

Acknowledgments

This paper was supported by the TRUMPET project, funded by the European Union under Grant Agreement No. 101070038.

References

[1] Loureiro Acuña, J., Martinez Luaña, X., Padín Torrente, H., Jiménez Balsa, G., García Pagán, C., Ortega Fernandez, I.: “Enhancing Privacy in Federated Learning: A Practical Assessment of Combined PETs in a Cross-Silo Setting”, in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*

A Review of "QKDNetSim+: Improvement of the Quantum Network Simulator for NS-3"

David Soler, Iván Cillero, Carlos Dafonte y Francisco Nóvoa
CITIC, Universidade da Coruña, A Coruña
Email: {david.soler,carlos.dafonte,fjnovoa}@udc.es

M. Fernández-Veiga y A. Fernández Vilas
atlanTTic, Universidade de Vigo, Vigo, Spain
Email: {mveiga,avilas}@det.uvigo.es

***Resumen*—En la actualidad se están desplegando las primeras redes de distribución cuántica de claves (QKD), pero el coste de implantación sigue siendo prohibitivo para la mayoría de los investigadores. Por ello, se necesitan simuladores de redes QKD realistas. El módulo *QKDNetSim* para el simulador de red NS-3 se centra en la representación de paquetes y la gestión de material de clave en una red QKD en la capa de aplicación. Sin embargo, algunos de sus componentes carecen de la profundidad que permitiría al simulador representar fielmente el comportamiento de una red cuántica real. En este trabajo, analizamos la arquitectura de QKDNetSim para identificar sus limitaciones, y presentamos una versión mejorada de QKDNetSim en la que se han modificado algunos de sus componentes para proporcionar un entorno de simulación más realista.**

***Index Terms*—Simulador de redes, QKD, QKDNetSim, NS-3**

Tipo de contribución: Investigación ya publicada. Trabajo original: *QKDNetSim+: Improvement of the quantum network simulator for NS-3* [1].

I. Introducción

Las tecnologías cuánticas han experimentado un avance significativo en los últimos años [2], [3], [4]. En concreto, los protocolos de distribución cuántica de claves (QKD) permiten a dos nodos acordar una clave a través de un canal cuántico, de tal forma que es imposible para un atacante obtener la clave sin ser detectado [5]. En la actualidad, el número de redes QKD implementadas es muy reducido debido al elevado coste del equipamiento necesario y a la falta de madurez de la tecnología. Por ello, los investigadores deben recurrir a simuladores que imiten el comportamiento de una red cuántica [6], [7].

Existe un módulo para NS-3 para la simulación de redes cuánticas, denominado QKDNetSim [8]. A diferencia de otros simuladores de redes cuánticas, QKDNetSim se centra en una perspectiva de red y en la gestión del material de clave generado mediante QKD. La estructura en capas de NS-3 también permite a QKDNetSim representar redes QKD a nivel de aplicación, en las que varios usuarios y programas comparten recursos QKD comunes, como canales cuánticos y almacenes de claves. Sin embargo, QKDNetsim contiene algunas limitaciones que afectan a su capacidad para representar fielmente el comportamiento de una red cuántica real.

Este trabajo es un resumen de *QKDNetSim+: Improvement of the Quantum Network Simulator for NS-3* [1], en el cual presentamos una versión mejorada del módulo de simulación de redes cuánticas QKDNetSim. Para ello, comenzaremos con un análisis de QKDNetSim: de qué elementos se compone y hasta qué punto imita el comportamiento de una red cuántica real, incluyendo las limitaciones y errores que impiden a QKDNetSim alcanzar los objetivos propuestos.

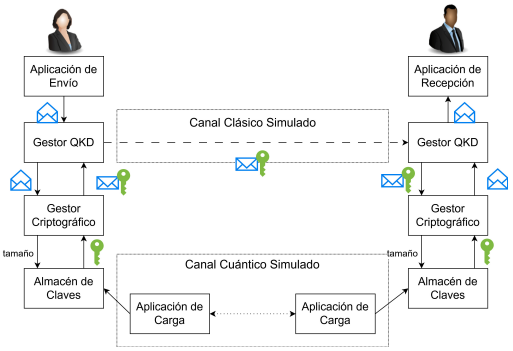


Figura 1. Arquitectura de QKDNetSim.

Nuestra implementación proporciona mejoras para superar las limitaciones mencionadas, manteniendo la arquitectura del módulo de simulación.

II. QKDNetSim

II-A. Arquitectura

El módulo QKDNetSim añade a NS-3 la posibilidad de crear canales cuánticos entre pares de nodos. Para ello, se añaden los siguientes componentes a cada par de nodos (tal como se muestra en la Figura 1:

- **Aplicaciones de Envío y Recepción**: simulan programas que crean paquetes para ser cifrados por el resto de componentes de QKDNetSim.
- **Gestor QKD**: componente central del módulo, que dirige la ejecución de los demás. El Gestor procesa los paquetes entrantes y salientes e identifica qué operaciones deben realizarse.
- **Gestor Criptográfico**: recibe peticiones del Gestor para cifrar o descifrar paquetes. Implementa múltiples algoritmos criptográficos y tiene acceso al Almacén de Claves.
- **Canal Cuántico Simulado**: imita el comportamiento de un canal cuántico. Cada uno de los nodos posee una *Aplicación de Carga*, que genera constantemente nuevo material de clave compartida.
- **Almacén de Claves**: Almacena el material de clave transmitido a través del canal cuántico para su uso futuro.

II-B. Limitaciones del módulo original

La implementación de estos componentes es poco realista en algunos aspectos, lo que impide a QKDNetSim ofrecer una simulación fiel. Destacamos las siguientes deficiencias de QKDNetSim:

1. **Todos los mensajes enviados a través del Canal Cuántico Simulado son idénticos**. El Canal Cuántico Simulado se encarga de la creación de nuevo material de clave entre dos nodos. Sin embargo, todos los mensajes intercambiados entre las Aplicaciones de Carga de nodos adyacentes contienen sólo una cadena de ≪0≫. La Aplicación de Carga del receptor no procesa los mensajes $ADDKEY$, puesto que ya sabe que la clave es una cadena de ≪0≫.
2. **Las claves no se almacenan**. El material de clave generado a través del Canal Cuántico Simulado no se almacena en ninguna parte. En su lugar, el Almacén de Claves sólo almacena un número que representa la *cantidad* de material de clave que debe tener. Cada vez que se añade o consume material de clave, este número se modifica en consecuencia. Dado que no hay material de clave real dentro de los Almacenes, tampoco existe ningún mecanismo que garantice, mantenga o recupere la sincronización entre los Almacenes de Claves de nodos conectados.
3. **Errores en el cifrado**. El Gestor Criptográfico asume incorrectamente que las claves se representan como lista de bytes, cuando en realidad se trata de bits. Cada vez que el simulador intenta cifrar un paquete, el programa aborta.

II-C. *Mejoras implementadas*

1. **Generación y Transmisión de Claves**: Las Aplicaciones de Carga procesan ahora adecuadamente los mensajes $ADDKEY$. Como en el módulo original, uno de los nodos se designa como *Primario* y el otro como *Secundario*: el nodo Primario se encarga de generar el material de clave y enviar los mensajes $ADDKEY$, mientras que el Secundario los recibe y procesa.
2. **Almacenamiento de Claves**: Los Almacenes de Claves se componen ahora de dos estructuras diferentes: el *almacenamiento en bruto* y la *Base de Datos de claves*. La segunda asigna a cada clave un índice y un tamaño específico. El proceso de obtener material de clave del almacenamiento en bruto e insertarlo en la Base de Datos de Claves con un índice se denomina *reservar*.
3. **Cifrado**: Los mensajes son cifrados por el Gestor Criptográfico y se les añade una cabecera que incluye la longitud del paquete y el índice de la clave utilizada.

II-D. *Ejecución de ejemplo*

```
QKDCrypto:ProcessOutgoingPacket(0x558a0f5a4d60, "SrcBuffer")
+20.256000000s 0 QKDBuffer:ReserveKeyMaterial(0x558a0f5a8460, " Reserved Key with ID ", 19, "of size", 720)
QKDCrypto:ProcessOutgoingPacket(0x558a0f5a4d60, "DstBuffer")
+20.256000000s 0 QKDBuffer:ReserveKeyMaterial(0x558a0f5a8770, " Reserved Key with ID ", 19, "of size", 720)
+20.256000000s 0 QKDBuffer:DeleteKeyID(0x558a0f5a8460, "KeyID is deleted from 'm_keys'", 19)
QKDCrypto:ProcessOutgoingPacket(0x558a0f5a4d60, "Encryption completed!")
QKDCrypto:ProcessOutgoingPacket(0x558a0f5a4d60, "Final outgoing packet from QCrypto:", "PacketID:", 340, "of
        size", 792, "MessageID:", 339, "QKDHeaderLength:", 792, "Encryption KeyID:", 19, "Auth KeyID: ", 20)

QKDCrypto:ProcessIncomingPacket(0x558a0f5a4d60, "PacketID:", 340, "PacketSize:", 792, 0x558a0f5a8770, "
        ChannelID:", 0)
+20.259270400s 1 QKDBuffer:DeleteKeyID(0x558a0f5a8770, "KeyID is deleted from 'm_keys'", 20)
+20.259270400s 1 QKDBuffer:FetchKeyByID(0x558a0f5a8770, "Fetching Key with ID: ", 19, " Value: ", "
        D08733E9CBFD7BF6[...]")
+20.259270400s 1 QKDBuffer:DeleteKeyID(0x558a0f5a8770, "KeyID is deleted from 'm_keys'", 19)
QKDCrypto:ProcessIncomingPacket(0x558a0f5a4d60, "Decrytion completed!", 0x558a0f6b10a0)

 [...]

+20.236610455s 0 QKDBuffer:AddKeyMaterial(0x558a0f5a8460, "m_Mcurrent:", 50960, "size:", 500, "key material :",
        220, 156, 42, 31, 103, 123, 152, 13,  [...],  17, 171, 245, 227, 25, 254, 225, 103)

QKDChargingApplication:HandleRead(0x55c03da58990, 0x55c03da66e80, "PACKETID: ", 325, " of size: ", 511)
QKDChargingApplication:ProcessIncomingPacket(0x55c03da58990, "Adding new key to DstBuffer")
+20.246034086s 1 QKDBuffer:AddKeyMaterial(0x558a0f5a8770, "m_Mcurrent:", 50960, "size:", 500, "key material :",
        220, 156, 42, 31, 103, 123, 152, 13,  [...],  17, 171, 245, 227, 25, 254, 225, 103)
```

Listado 1. Ejecución de ejemplo

El Listado 1 muestra una ejecución de ejemplo del simulador. En la primera sección se reserva material de clave en ambos Almacenes de Claves para cifrar un mensaje. El algoritmo de cifrado es One-Time-Pad, lo que determina el tamaño de la clave a reservar. A continuación, el receptor procesa este paquete y obtiene la clave de descifrado de su Almacén de Claves utilizando el índice (presente en la cabecera del mensaje). La última sección muestra un intercambio entre las Aplicaciones de Carga para añadir 500 bytes de nuevo material a los Almacenes de Claves de ambos nodos.

III. Conclusión

En este trabajo hemos analizado el módulo de simulación de redes cuánticas QKDNetSim de NS-3, describiendo sus principales componentes e identificando sus limitaciones y su impacto en la calidad de la simulación de una red QKD. Nuestra implementación de un QKDNetSim mejorado mantiene la estructura general del módulo pero supera las limitaciones de sus componentes. Nuestra versión proporciona una representación fiel de las redes QKD, lo que permite a los usuarios comprender mejor las operaciones relacionadas con la gestión de claves en la capa de aplicación.

Reconocimiento

Este trabajo está financiado por el Plan Complementario de Comunicaciones Cuánticas, Ministerio de Ciencia e Innovación (MICINN), Plan de Recuperación NextGeneration, Unión Europea (PRTR-C17.I1, CITIC Ref. 305.2022), y la Xunta de Galicia (Agencia Gallega de Innovación, GAIN, CITIC Ref. 306.2022) D.S. reconoce apoyo de la Xunta de Galicia - Beca [ED481A-2023-219].

Este trabajo es parte de los proyectos TED2021-130369B-C31 and TED2021-130492B-C21, financiados por MCIN/AEI/10.13039/501100011033 y "ERDF A way of making Europe".

Referencias

[1] D. Soler, I. Cillero, C. Dafonte, M. Fernández-Veiga, A. F. Vilas, F. J. Nóvoa, Qkdnetsim+: Improvement of the quantum network simulator for ns-3, SoftwareX 26 (2024) 101685. doi:https://doi.org/10.1016/j.softx.2024.101685.

[2] J. Illiano, M. Caleffi, A. Manzalini, A. S. Cacciapuoti, Quantum internet protocol stack: A comprehensive survey, Computer Networks 213 (2022) 109092. doi:10.1016/j.comnet.2022.109092.

[3] A. Singh, K. Dev, H. Siljak, H. D. Joshi, M. Magarini, Quantum internet—applications, functionalities, enabling technologies, challenges, and research directions, IEEE Communications Surveys and Tutorials 23 (4) (2021) 2218–2247. doi:10.1109/COMST.2021.3109944.

[4] Y. Cao, Y. Zhao, Q. Wang, J. Zhang, S. X. Ng, L. Hanzo, The evolution of quantum key distribution networks: On the road to the qinternet, IEEE Communications Surveys & Tutorials 24 (2) (2022) 839–894. doi:10.1109/COMST.2022.3144219.

[5] R. Alléaume, C. Branciard, J. Bouda, T. Debuisschert, M. Dianati, N. Gisin, M. Godfrey, P. Grangier, T. Länger, N. Lütkenhaus, et al., Using quantum key distribution for cryptographic purposes: a survey, Theoretical Computer Science 560 (2014) 62–81. doi:10.1016/j.tcs.2014.09.018.

[6] T. Coopmans, R. Knegjens, A. Dahlberg, D. Maier, L. Nijsten, J. de Oliveira Filho, M. Papendrecht, J. Rabbie, F. Rozpkedek, M. Skrzypczyk, et al., Netsquid, a network simulator for quantum information using discrete events, Communications Physics 4 (1) (2021) 164. doi:10.1038/s42005-021-00647-8.

[7] S. Diadamo, J. Notzel, B. Zanger, M. M. Bese, QuNetSim: A software framework for quantum networks, IEEE Transactions on Quantum Engineering 2 (2021) 1–12. doi:10.1109/tqe.2021.3092395.

[8] M. Mehic, O. Maurhart, S. Rass, M. Voznak, Implementation of quantum key distribution network simulation module in the network simulator ns-3, Quantum Information Processing 16 (2017) 1–23.

Deploying a Hybrid Quantum-Safe VPN for Secure Banking Communications

Raúl Díaz Bonete
Indra Sistemas S.A.
Madrid, Spain
rdbonete@indra.es

Marta Irene García Cid
Indra Sistemas S.A.
Madrid, Spain
migarcia@indra.es

David Domingo Martín
Indra Sistemas S.A.
Madrid, Spain
ddomingo@indra.es

***Abstract*—This work addresses the security of backbone communications in the banking sector against emerging threats from quantum computing. It proposes a hybrid solution to replace classic asymmetric key exchange in IPsec-based VPNs with cryptographic mechanisms resistant to quantum attacks. The proposed architecture combines three key negotiation schemes: classical asymmetric cryptography (ECDH P-384), post-quantum cryptography (PQC ML-KEM-768), and quantum key distribution (QKD BB84 & GMCS), which are combined through a quantum safe hybridization function, in order to obtain a final secure shared secret. The architecture is designed to enable hybrid key negotiation over a 60 km distance. Three nodes will be deployed, allowing the establishment of IPsec tunnels between any of them. If a tunnel needs to be established between the most distant nodes, the intermediate node will act as a key relay, to solve the distance problem posed by signal losses in optical fiber. Technologies for both discrete-variable and continuous-variable optical links will be integrated. The proposed architecture uses software-defined networking (SDN) to efficiently manage the distribution of quantum keys between data centers and to secure the exchange of sensitive business information.**

***Index Terms*—Banking Security, IPsec VPN, Post-Quantum Cryptography (PQC), Quantum Key Distribution (QKD), Key Hybridization, Software-Defined Networking (SDN)**

Type of contribution: *Ongoing research*

I. Introduction

Network communication security is one of the most critical concerns in today's technological landscape. In this context, one of the critical industries that requires greater security is the banking sector, whose activity is essential for economic development and financial stability. Modern banking systems face multiple threats and external attacks, which makes it imperative to ensure security for both bank premises and customer services. A key objective is to predict which technologies will emerge as effective solutions to these challenges [1]. In this context, quantum key distribution (QKD) and post-quantum cryptography (PQC) stand out as advanced technologies in the realm of communication security.

One of the pillars of the banking infrastructure are the data centers and links that connect them. Current security relies on protocols like MACsec, IPsec, and TLS. However, the rise of quantum computing poses a threat to the classical asymmetric cryptosystems used in these protocols.

The objective of this work is to replace traditional asymmetric key exchange in IPsec based VPN with a quantum-resistant negotiation. To achieve this purpose, a hybrid approach is employed, combining classical, PQC, and QKD keys for enhanced security and resilience, against current and future cryptographic threats.

II. IPSec protocol

Internet Protocol Security (IPsec) is the most commonly used network layer security control to protect communications. It is designed to ensure interoperability, and depending on its configuration, it can provide integrity, authentication, confidentiality, traffic analysis or replay protection.

To establish a secure and authenticated channel, the Internet Key Exchange (IKE) protocol is used. IKE is responsible for negotiating cryptographic parameters, authenticating peers, and establishing Security Associations (SAs). A SA is uniquely identified by a Security Parameter Index (SPI), an IPv4 or IPv6 destination address, and a security protocol (AH or ESP) identifier. SAs are security agreements that may include the following parameters: symmetric encryption protocol, authentication algorithm, key negotiation algorithm, security mode, lifetime and a packet antireplay mechanism [2], [3].

Traditionally, the key negotiation algorithm relies on asymmetric cryptosystems such as RSA (Rivest, Shamir and Adleman), DH (Diffie-Hellman) and ECDH (Elliptic Curve Diffie-Hellman). Their security is based on the hardness of factoring large integers into their prime factors, the DLP (discrete logarithm problem) and the ECDLP (Elliptic curve discrete logarithm problem) respectively. In general, signature algorithms rely on key exchange mechanisms based on asymmetric cryptography, such as DH or RSA combined with Message Authentication Codes (MACs) like HMAC or KMAC, or alternatively on purely asymmetric signature schemes such as ECDSA (Elliptic Curve Digital Signature Algorithm). However, classical asymmetric cryptography is vulnerable to attacks based on quantum computing, calling into question the security foundation in messaging systems, password management, or financial transactions, among others.

III. Quantum threats & alternatives to classical asymmetric cryptography

Quantum computers use the principles of quantum mechanics to solve complex problems. They can break classical asymmetric encryption systems such as RSA, DH, and ECDH through quantum algorithms. Shor's Algorithm (1994) [5] can factor large numbers, breaking RSA. It can be adapted to solve DLP and ECDLP, breaking DH and ECDH respectively. Regev's Algorithm (2024) [6] also solves integer factorization, offering another way to break RSA. Consequently, asymmetric cryptography-based signing schemes are also broken, such as the Elliptic Curve Digital Signature Algorithm (ECDSA).

For this, various solutions are proposed. On the one hand, National Institute of Standards and Technology (NIST) supports the migration to PQC. PQC refers to algorithms designed to resist quantum-based attacks, founded on computationally complex mathematical problems for both classical and quantum computers. The main approaches are the following: **Lattice-based** schemes rely on problems like Module Learning With Errors (MLWE), as used in the Module Lattice Key Encapsulation Mechanism (ML-KEM) defined in FIPS 203 [7] and the Module Lattice Digital Signature Algorithm (ML-DSA) in FIPS 204 [8]; **Code-based** schemes, such as Classic McEliece and HQU, are based on the difficulty of decoding random linear codes; and **Symmetric-based** schemes include stateless hash-based digital signatures like SLH-DSA, defined in FIPS 205 [9].

On the other hand, quantum key distribution (QKD) emerges. It can be proven to be Information-Theoretic Secure (ITS), meaning that its security is not reliant on computational assumptions. QKD ensures that two parties securely share a random key by relying on quantum principles like the no-cloning theorem or Heisenberg's uncertainty principle. These principles prevent eavesdroppers from intercepting the key transmitted [10].

IV. Quantum Safe Hybridization

It is clear that the development of increasingly powerful quantum computers poses a serious threat to the security of classical asymmetric cryptography. However, it remains uncertain when the computational power of quantum computers will be sufficient to completely render it ineffective. On the other side, PQC systems are emerging as resistant to quantum attacks; nevertheless, since their security relies on complex mathematical problems, it is unclear whether, as attack methods and techniques evolve, they will remain secure or eventually be compromised. Finally, Quantum Key Distribution (QKD) can be proven mathematically secure, but its physical implementation presents significant challenges, such as the need for single-photon sources and detectors, optic fibre losses, among others [11].

The concept of combining multiple primitives of the same type to ensure that the resulting scheme remains secure as long as at least one of the components remains secure dates back to the work of Even and Goldreich [12] [13]. Thus, combining all three schemes (asymmetric, PQC and QKD) presents a strong alternative for transitioning to secure cryptography. If any of the hybridized mechanisms are compromised, the security remains intact, as the others provide protection. The hybridization scheme used in the banking use case will consist of a combination of an ECDH key, a ML-KEM key, and a QKD key as defined in ETSI TS 103 744 [14].

A. Elliptic Curve Diffie-Hellman

The first of the three key agreement schemes is ECDH. A shared secret k is computed between two entities over the same elliptic curve domain parameters.

The key agreement scheme, for a given set of elliptic curve domain parameters, consists of a pair of algorithms:

- **ECKeyGen()**: This is a probabilistic algorithm that returns a private and public key pair (d, Q).
- **ECDH**(d_A, Q_B): A deterministic algorithm that takes d_A (entity A's private key) and Q_B (entity B's public key) as input and computes a shared secret k, or $\perp$ in case of an error. Similarly, party B computes the same shared secret by executing $k = \text{ECDH}(d_B, Q_A)$.

In our case, we will be using NIST curve P-384 defined in clause 3.2.1.4 of NIST SP 800-56Ar3 [4].

B. Key Encapsulation Mechanism

The second type of key agreement mechanism implemented is a Key Encapsulation Mechanism (KEM), which is defined as a set of algorithms (KEMKeyGen, Encaps, Decaps) associated with a key space K.

- **KEMKeyGen()**: A probabilistic algorithm that returns a key pair (sk_A, P_A).
- **Encaps**(P_A): A probabilistic algorithm that takes a public key P_A and outputs a cryptographic key k and an encapsulation C.
- **Decaps**(sk_A, C): A deterministic algorithm that takes a private key sk and an encapsulation C to return a cryptographic key k.

The ML-KEM that will be used is ML-KEM-768 [15].

C. Quantum Key Distribution

The security of Quantum Key Distribution (QKD) relies on quantum mechanics to securely generate and share secret keys between two parties using quantum states, typically represented as qubits and physically implemented with photons transmitted through media like optical fibers or free space. The ETSI GS QKD 014 standard [17] defines a REST API that enables a Secure Application Entity (SAE) to access QKD keys stored in a Key Management Entity (KME), promoting interoperability across different vendors. The initiating SAE is called the master SAE and may connect to one or more KMEs. The **enc_keys()** method allows the master SAE to request a new QKD key from the master KME, which returns the key material and an associated key ID. This ID is then communicated to the slave SAE, which can use the **dec_keys(id)** method to retrieve the corresponding key material from its own KME using the key ID.

D. Quantum Safe Cascade Key Derivation Function

The selected hybridization function utilizes the shared secrets, public and private cryptographic material generated during the exchanges. It follows a cascade scheme, performing as many rounds as the number of key exchange algorithm types used. It is based on the following functions.

- **KDF.** A Key Derivation Function (KDF) is a fundamental component in cryptographic systems. Its purpose is to take an initial keying material as input and derive one or more secret keys from it in a cryptographically secure manner. In this implementation, HMAC-based Extract-and-Expand Key Derivation Function (HKDF) is used, which is specified in IETF RFC 5869 [18].
- **PRF.** A Pseudo-Random Function (PRF) is designed to generate an output from a given random, or secret, input. The principal property of a PRF is that its output is computationally indistinguishable from a truly random sequence. In this implementation, Hash-based Message

Authentication Code (HMAC) function is used, as defined in IETF RFC 2104 [19].

The input parameters are as follows: psk is a digest_len-length secret key obtained via quantum key distribution; $(k_1, k_2, \ldots, k_n)$ is a sequence of shared secrets established through key exchange; MA_i and MB_i are exchanged messages containing the public elements required for each key exchange ($i = 1, \ldots, n$); ctx_i denotes the context for establishing each secret; lbl_i is a label octet string that scopes the usage within an application or key exchange instance; and len_i indicates the desired length for intermediate and final keys for each i.

Algorithm 1 Quantum Safe KDF

```
chain_s_0 ← psk.
For i ← 1, ..., n:
    round_s_i ← PRF(chain_s_{i-1}, k_i, MA_i, MB_i).
    kdf_len ← digest_len + len_i.
    chain_s_i , key_mat_i ←
            KDF(round_s_i, lbl_i, ctx_i, kdf_len).
Return chain_s_i and key_mat_i.
```

Finally, the output of the quantum safe hybridization function is a sequence of intermediate and final keys key_mat_i and a sequence of chain secrets $chain_s_i$.The final quantum-safe key is key_mat_n.

V. Banking Use Case

This use case involves establishing an IPsec VPN tunnel to connect data centers that support the core communications of a bank. Given the critical nature of these communications, it is essential to ensure end-to-end encryption (E2EE), regardless of distance. The range of commercial QKD systems is typically limited to 100 kilometers over optical fibers [20] (free-space transmission is not considered in the use case presented in this work).

The two main actors in this scenario are the Bank and the Carrier. The Bank must deploy devices in each of its data centers capable of establishing VPN tunnels to route all business-related traffic. The Bank must be able to perform ECDH P-384 and ML-KEM-768 key exchanges, which will then be hybridized with the QKD keys provided by the Carrier.

The Carrier is responsible for providing end-to-end quantum telecommunications between the data centers. Due to the no-cloning theorem, a quantum state cannot be copied. To overcome signal losses, key forwarding is necessary. As a result, the Carrier will participate in the QKD exchange by incorporating the necessary nodes along the intended communication path. QKD serves here as a last resort for crypto-agility, meaning that in the unlikely event that both the ECDH P-384 and ML-KEM-768 schemes are compromised, QKD will keep the communications secure. Additionally, the bank's trust in the Carrier is limited, as the Carrier is not responsible for obtaining all the keys that will secure the communications.

The deployment of QKD-based networks presents challenges beyond scalability. Infrastructure costs are high, as QKD systems require specialized hardware and significant investment, making large-scale deployment financially demanding.

VI. Physical Architeture

The physical architecture will be deployed across three independent buildings as shown in Fig. 1. For the quantum channels, dark fiber optic links a 24.3 km fiber with an attenuation of less than 8 dB connects Building A and Building B. The optical fiber connecting Building B to Building C has a length of 41.8 km and an attenuation of less than 20 dB. Different QKD device technologies are employed.

The first quantum key distribution devices are based on Discrete Variable QKD (DV-QKD). The standard BB84 decoy state protocol is used, which includes four-state preparation and measurement modules, respectively. The physical encoding is done in the time-bin basis, meaning that one degree of freedom is the time of arrival of the photon (early/late), and the second is the phase relationship between them.

The second installed QKD modules are based in Continuous Variable QKD (CV-QKD). This devices implement the GMCS protocol, in which both the X and P quadratures are independent and modulated according to a zero-centered Gaussian random distribution. The post-processing includes downsampling, quantum state phase recovery, pattern synchronization, and the estimation of various parameters such as excess noise, channel transmittance, and secret key rate. This process is computationally expensive, which is why these devices use an external dedicated server (CV-Server) to perform these post-processing tasks.

CV-QKD has recently gained a lot of interest due to its ability to be implemented using standard telecom components that function at room temperature. However, these devices usually have a shorter range than discrete-variable ones, and their security proof is more challenging as in [21], [22].

All the infrastructure necessary for key management will be deployed on three independent servers. These will be equipped with HSMs (Hardware Security Modules) as in Fig. 1. HSMs are part of the Tamper Resistant Security Module (TRSM) or Secure Cryptographic Device (SCD) family. These are physically secure and/or tamper-responsive devices designed for cryptographic processing [23]. In the proposed architecture, they will be responsible for storing the essential keys and performing all cryptographic operations, which we have already described in Section IV.

The encryptors that will create the VPN tunnel will be next-generation firewalls (NGFW), which will be responsible for establishing the IPsec tunnel with a bandwidth of at least 10 Gbps using the keys provided by the KMEs. This tunnel will be responsible for supporting core banking communications, transmitting essential information for business development.

All components within each building will be connected through a switch. Equipment configuration and infrastructure management can be done either via LAN from within the hosting building or remotely through a VPN established by the NGFWs.

VII. Logical architecture

Software-Defined Quantum Key Distribution (SD-QKD) network combines QKD with Software-Defined Networking (SDN) to make secure key exchange more efficient, flexible, and automated. A central or distributed controller manages KMEs, enabling direct quantum links and multi-hop virtual

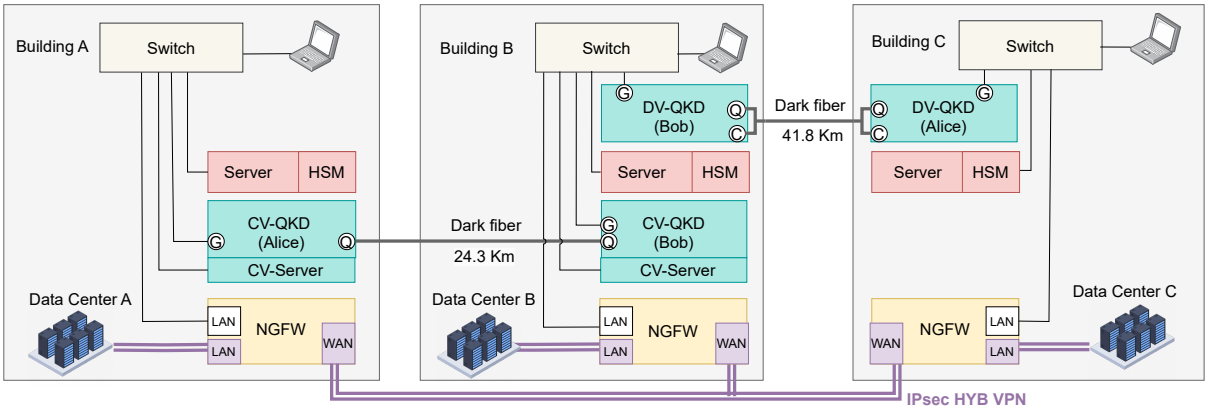


Figure 1. Physical System Architecture Diagram

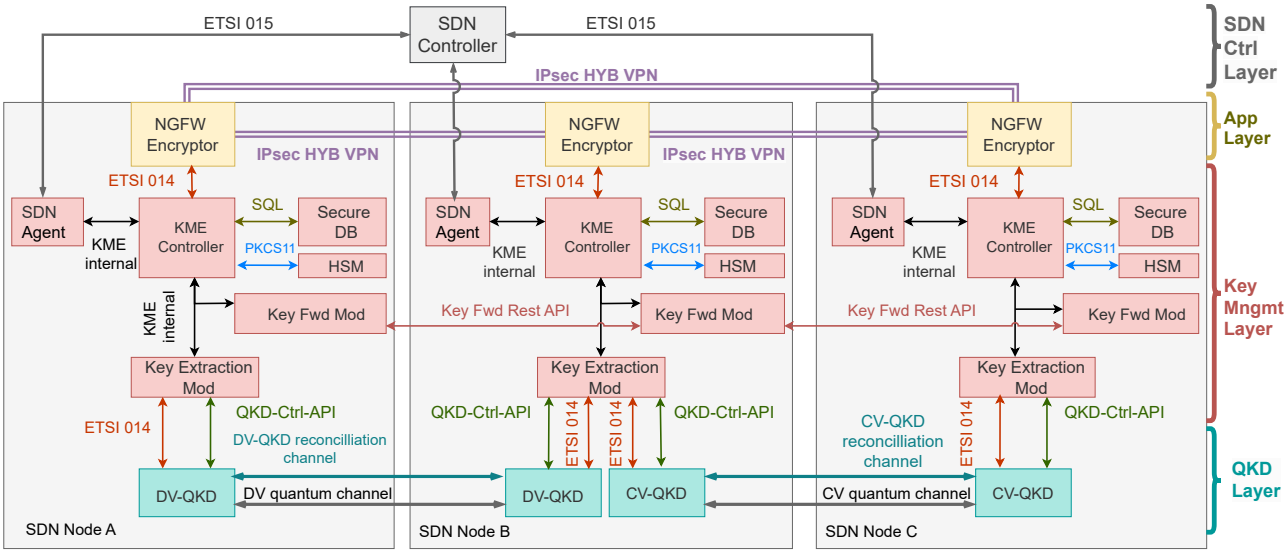


Figure 2. Logical System Architecture Diagram

connections based on network needs. This approach simplifies integration into modern networks, making them more adaptable and scalable for real-world cybersecurity challenges.

The SDN controller is responsible for extracting network metrics to enhance decision-making and path selection. It also manages the negotiation of cryptographic keys, whether hybrid or QKD, for each physical link that constitutes a virtual link. Once all necessary keys are available on the physical links, the controller orchestrates key relaying to extend them to the virtual link.

An SD-QKD node aggregates one or more QKD modules and interfaces with an SDN controller to enable dynamic key management and network optimization. It securely manages key associations via direct quantum channels or multi-hop links while exposing QKD system information and classical channel requirements to the controller. The node ensures seamless key retrieval through an API and supports telemetry, discovery, and control functions. By abstracting QKD systems within a secure location, it facilitates end-to-end key associations without compromising security, as encryption keys are never transmitted to the controller. Our infrastructure will be based on an SDN network structured into multiple layers, according to Fig. 2. This type of architecture is defined in the ETSI GS QKD 015 [24] and ITU Y.3803 [25] standards.

The application layer includes cryptographic application entities that consume shared secret key pairs provided by the QKD-SDN. In the present implementation, this layer consists of NGFWs, which will utilize the keys provided through the ETSI 014 interface, as in IV-C.

The key management layer handles the reception, storage, and forwarding of keys generated by QKD devices for cryptographic applications. It includes an SDN agent that standardizes protocols between the SDN network and the KME. The KME controller interconnects this layer within the SDN node. The Key Extraction Module retrieves QKD keys from QKD devices in the node, aiming to extract keys at the Key Rate specified by the QKD module, minimizing the time keys remain in non-secure storage. This rate is determined through the QKD module's proprietary control interface.

Keys are securely stored, wrapped by a Key Encryption Key (KEK), which is kept in the HSM. Additionally, the entire database (DB) is encrypted with a DB key, also stored in the

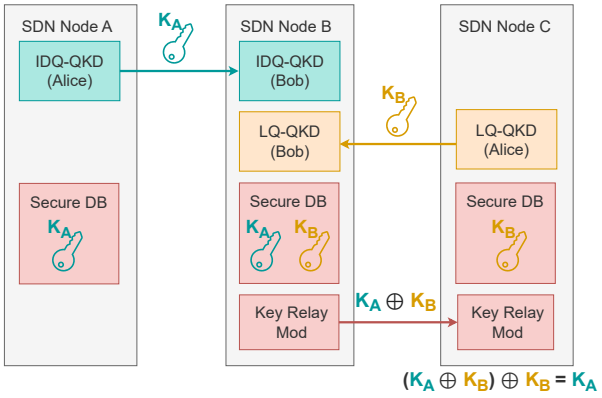


Figure 3. Key Relay Algorithm.

HSM.

The Key Forwarding Module is responsible for managing the transmission and reception of forwarded keys, handling the public elements required for key agreements, while all cryptographic operations involving these elements are performed within the HSM. The forwarding method is based on encrypting the key to be relayed using a One-Time Pad (OTP) with another key stored in the link where the key is to be extended. This Vernam encryption process also occurs within the HSM. In Fig. 3, it is shown how the key K_A can be securely extended to the virtual link (SDN Node A - SDN Node C) using the key K_B.

VIII. Conclusions

Network communication security is essential for the banking sector, especially in light of the threats that quantum computing poses to traditional cryptographic systems. This work presents a hybrid solution that combines classical, post-quantum, and quantum cryptography to secure communications between banking data centers.

The proposed architecture demonstrates that it is possible to implement a quantum-resistant security system in real banking infrastructures by utilizing a software-defined network that efficiently manages key distribution across multiple nodes. This hybrid approach enables a gradual transition toward fully quantum-resistant systems, ensuring that even if one cryptographic mechanism is compromised, the overall system remains secure. The architecture presented in this work is currently under deployment. Once completed, the use case will be executed.

Acknowledgments

This project has received funding from the Digital Europe Programme (EuroQCI Spain) under grant agreement Nº 101091638. The authors would like to acknowledge Jaime Gómez García for his valuable inspiration and insightful contributions to the development of this work.

References

[1] M. N. Kondratyeva, D. D. Svirina, and A. I. Tsvetkov: The role of information technologies in ensuring banking security, in *IOP Conference Series: Materials Science and Engineering*, vol. 1047, no. 1, 2021.

[2] S. E. Frankel, K. Kent, R. Lewkowski, A. D. Orebaugh, R. W. Ritchey, and S. R. Sharma, SP 800-77. Guide to IPsec VPNs, National Institute of Standards and Technology (NIST), 2005.

[3] C. Kaufman, P. Hoffman, Y. Nir, P. Eronen, and T. Kivinen, RFC 7296: Internet Key Exchange Protocol Version 2 (IKEv2), Internet Engineering Task Force (IETF), 2014. [Online]. Available: https://www.rfc-editor.org/info/rfc7296

[4] E. Barker, L. Chen, A. Roginsky, A. Vassilev, and R. Davis, Recommendation for Pair-Wise Key-Establishment Schemes Using Discrete Logarithm Cryptography, *NIST Special Publication 800-56A Revision 3*, National Institute of Standards and Technology, 2018. [Online]. Available: https://csrc.nist.gov/pubs/sp/800/56/a/r3/final

[5] P. W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pp. 124-134, 1994.

[6] O. Regev, An efficient quantum factoring algorithm, *Journal of the ACM*, vol. 72, no. 1, pp. 1-13, 2025.

[7] National Institute of Standards and Technology (NIST), FIPS 203: Module-Lattice-Based Key-Encapsulation Mechanism Standard, Department of Commerce, Washington, D.C., 2024. [Online]. Available: https://doi.org/10.6028/NIST.FIPS.203

[8] National Institute of Standards and Technology (NIST), FIPS 204: Module-Lattice-Based Digital Signature Standard, Department of Commerce, Washington, D.C., 2024. [Online]. Available: https://doi.org/10.6028/NIST.FIPS.204

[9] National Institute of Standards and Technology (NIST), FIPS 205: Stateless Hash-Based Digital Signature Standard, Department of Commerce, Washington, D.C., 2024. [Online]. Available: https://doi.org/10.6028/NIST.FIPS.205

[10] Y. Cao, Y. Zhao, Q. Wang, J. Zhang, S. X. Ng, and L. Hanzo, The Evolution of Quantum Key Distribution Networks: On the Road to the Qinternet, *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 839-894, Second quarter 2022.

[11] C. Marquardt, U. Seyfarth, *et al.*, Implementation Attacks against QKD Systems, 1.0, 21 Nov. 2023.

[12] S. Even and O. Goldreich, On the power of cascade ciphers, *ACM Transactions on Computer Systems*, vol. 3, no. 2, pp. 108–116, May 1985.

[13] S. Ricci, et al., Hybrid keys in practice: combining classical, quantum and post-quantum cryptography, *IEEE Access*, vol. 12, pp. 23206–23219, 2024.

[14] ETSI, ETSI TS 103 744 V1.1.1 (2020-12) CYBER; Quantum-safe Hybrid Key Exchanges, *ETSI Technical Specification*, 2020.

[15] R. Avanzi, J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, ... and D. Stehlé, CRYSTALS-Kyber algorithm specifications and supporting documentation, *NIST PQC Round*, vol. 2, no. 4, pp. 1-43, 2019.

[16] I. B. Djordjevic, Physical-layer security and quantum key distribution, *2019*, pp. 211-265.

[17] ETSI, Quantum Key Distribution (QKD); Protocol and data format of REST-based key delivery API, ETSI GS QKD 014 V1.1.1, Feb. 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/QKD/001_099/014/01.01.01_60/gs_qkd014v010101p.pdf

[18] H. Krawczyk and P. Eronen, HMAC-based Extract-and-Expand Key Derivation Function (HKDF), *Request for Comments (RFC 5869)*, 2010. [Online]. Available: https://www.rfc-editor.org/info/rfc5869

[19] H. Krawczyk, M. Bellare, and R. Canetti, HMAC: Keyed-Hashing for Message Authentication, *Request for Comments (RFC 2104)*, 1997. [Online]. Available: https://www.rfc-editor.org/info/rfc2104

[20] B. Huttner, R. Alléaume, E. Diamanti, F. Fröwis, P. Grangier, H. Hübel, ... and H. Zbinden, Long-range QKD without trusted nodes is not possible with current technology, *npj Quantum Information*, vol. 8, no. 1, p. 108, 2022.

[21] E. Diamanti and A. Leverrier, Distributing secret keys with quantum continuous variables: principle, security and implementations, *Entropy*, vol. 17, no. 9, pp. 6072-6092, 2015.

[22] N. Jain, et al., Practical continuous-variable quantum key distribution with composable security, *Nature Communications*, vol. 13, no. 1, p. 4740, 2022.

[23] S. Mavrovouniotis and M. Ganley, Hardware security modules, in *Secure Smart Embedded Devices, Platforms and Applications*, New York, NY: Springer New York, 2013, pp. 383-405.

[24] ETSI GS QKD 015, Quantum Key Distribution; Control Interface for Software Defined Networks, Apr. 2022. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/QKD/001_099/015/02.01.01_60/gs_QKD015v020101p.pdf

[25] ITU-T Y.3803, Quantum Key Distribution Networks – Key Management, Dec. 2020. [Online]. Available: https://www.itu.int/rec/T-REC-Y.3803-202012-I/en

Análisis de Privacidad en Aplicaciones Móviles y Redes Sociales mediante Técnicas de IA

Daniel Blanco-Aza*, Antonio Robles-Gómez*, Rafael Pastor-Vargas*, Llanos Tobarra*,
Pedro Vidal-Balboa†, Mariano Méndez-Suárez‡
*Departamento de Sistemas de Comunicación y Control, ETSI Informática,
Universidad Nacional de Educación a Distancia (UNED), Madrid, España
Email: {d.blanco,arobles,rpastor,llanos}@scc.uned.es
†Consorcio CA UNED en Ponferrada / INTECCA,
Email: pvidal@intecca.uned.es
‡ESIC University, ESIC Business & Marketing School,
Email: mariano.mendez@esic.university

***Resumen*—En el contexto actual de las aplicaciones móviles y las redes sociales, la privacidad se ha convertido en una preocupación clave debido al amplio manejo y recopilación de datos personales. Por ello, este trabajo ofrece una revisión exhaustiva del estado del arte sobre el análisis automatizado de riesgos de privacidad en estos entornos. Se examinan diferentes enfoques, herramientas y marcos que emplean técnicas de aprendizaje automático y procesamiento del lenguaje natural para evaluar el cumplimiento de normativas como el RGPD. A través de un análisis sistemático basado en la metodología PRISMA, se identifican avances significativos en la aplicación de métodos automatizados para la evaluación de riesgos de privacidad.**

***Index Terms*—Riesgos de Privacidad, Aplicaciones Móviles, Inteligencia Artificial (IA), PRISMA, Protección de Datos.**

Tipo de contribución: *Investigación ya publicada* [1]

I. Introducción

La expansión de aplicaciones móviles y redes sociales ha supuesto un cambio profundo en la forma de comunicarnos y gestionar datos personales, intensificando al mismo tiempo las inquietudes en torno a la privacidad. El Reglamento General de Protección de Datos (RGPD) [2] exige transparencia y salvaguarda de la información, pero la diversidad de plataformas y la complejidad de los modelos de negocio dificultan su cumplimiento efectivo. En este escenario, surgen metodologías basadas en técnicas de procesamiento de lenguaje natural (Natural Language Processing, NLP) y técnicas de Inteligencia Artificial (IA) mediante el aprendizaje automático (Machine Learning, ML) para identificar prácticas inadecuadas de recolección, uso y difusión de datos, entre otras acciones maliciosas.

Muchos usuarios desconocen los mecanismos de recolección de información en estas plataformas; en particular, poblaciones vulnerables como menores de edad enfrentan riesgos agravados [3]. Con el fin de abordar esta problemática, se precisa investigar soluciones que automaticen el análisis de políticas de privacidad y permisos, y que provean un marco integral de evaluación de riesgos.

El estudio original del estado del arte fue presentado originalmente en *The Fourth Intelligent Cybersecurity Conference (ICSC2024)* [1], en el cual se repasaba con más detalle los distintos análisis de riesgos de privacidad en aplicaciones móviles y redes sociales aplicando técnicas de IA. Se está desarrollando una aplicación basada en dicho estudio que sirva de semáforo para detectar estos riesgos de privacidad.

II. Enfoques de Análisis de Riesgos de Privacidad

Se llevó a cabo el estudio del estado del arte en cuanto a herramientas sobre los distintos enfoques de análisis de riesgos de privacidad. En primer lugar, AC-Net [4] verifica la coherencia entre los permisos declarados por una app Android y su descripción en la tienda, usando modelos de NLP y redes neuronales para detectar inconsistencias. ATLAS [5] compara políticas de privacidad con las etiquetas *privacy labels* de apps iOS mediante un clasificador que predice las etiquetas a partir del contenido de la política, señalando discrepancias.

Por su parte, ATPChecker [6] examina si las librerías de terceros (TPLs) en Android cumplen regulaciones de privacidad. Su enfoque combina la recolección sistemática de políticas de TPL con análisis estático del código para rastrear flujo de datos personales. Cualquier discordancia entre lo declarado y lo observado se etiqueta como posible falta de conformidad. iABC-AL [7] clasifica apps iOS como benignas, sospechosas o maliciosas, asociando el uso de permisos con su categoría mediante aprendizaje activo para detectar comportamientos anómalos. PermPress [8] investiga la integridad de los permisos *peligrosos* en aplicaciones móviles en Android al comparar su mención en la política de privacidad con los permisos detectados en el binario.

Además de ello, PolicyChecker [9] revisa la *completitud* de políticas de privacidad según el GDPR, identificando omisiones en cláusulas obligatorias o condicionales. Además, PolicyLint [10] examina políticas de privacidad de apps en Google Play para identificar contradicciones internas, como declaraciones encontradas en distintos apartados que se anulan mutuamente. Estas ambigüedades pueden confundir a los usuarios y vulnerar la claridad exigida por la regulación.

En este sentido, PRADroid [11] combina clasificadores *ensemble* y análisis de flujo de información para evaluar el riesgo de fugas de datos en Android, asignando puntuaciones de exposición de privacidad. PriVot [12] monitoriza el tráfico de red en dispositivos Android para comparar datos recolectados con lo declarado en la política de privacidad, ofreciendo resúmenes claros al usuario. Finalmente, XPOChecker [13] analiza la *sobreexposición* de datos personales en apps mediante pruebas estáticas y dinámicas, detectando si la información privada se comparte indebidamente.

El análisis realizado evidencia varias deficiencias en las

herramientas actuales de evaluación de riesgos de privacidad. Ninguna de ellas implementa un sistema de puntuación detallado que permita evaluar la gravedad de cada caso de manera estructurada. Además, la falta de visualizaciones gráficas dificulta la interpretación de los riesgos y, por ende, la toma de decisiones informadas por parte de los usuarios. Por último, no existe una solución unificada capaz de cuantificar integralmente todos los riesgos de privacidad y proporcionar una evaluación detallada del nivel de exposición de los datos. Estas carencias revelan una brecha significativa en el panorama actual, ya que ninguna herramienta ofrece una solución completa que aborde estos aspectos de manera conjunta y efectiva.

III. Conclusiones y Líneas Futuras

Esta investigación ha identificado diversas metodologías, herramientas y marcos que permiten realizar análisis de privacidad automatizados con técnicas de IA. Sin embargo, aún persisten desafíos, especialmente la falta de una herramienta unificada que permita cuantificar y analizar detalladamente todos los riesgos de privacidad de manera simultánea.

Destaca la necesidad de un enfoque integral para la evaluación de riesgos de privacidad, con el objetivo de desarrollar un marco detallado, cuantificable y fácil de usar en aplicaciones móviles y redes sociales. Esto permitiría a los usuarios proteger mejor sus datos personales y garantizar el cumplimiento normativo, promoviendo un entorno digital más seguro.

Como trabajo futuro, estamos desarrollando un marco integral de evaluación de riesgos de privacidad adaptado a aplicaciones móviles y redes sociales, además de una aplicación móvil que ofrezca soporte al usuario, y que abarque un amplio espectro de problemas de privacidad, mejorando su precisión, escalabilidad y usabilidad. Además, se está explorando su aplicación en otros ámbitos, como los videojuegos y las aplicaciones de salud, para fortalecer la protección de datos en distintos sectores.

Agradecimientos

Los autores desean agradecer el apoyo del Gobierno de España (a través del Instituto Nacional de Ciberseguridad de España, INCIBE), con el Proyecto de Investigación Estratégico “Análisis de aplicaciones móviles desde la perspectiva de la protección de datos: Ciberprotección y Ciberriesgos de la información ciudadana”, así como la Cátedra Internacional “Smart Rural IoT y Entornos Seguros”, en el contexto del Plan de Recuperación, Transformación y Resiliencia financiado por la Unión Europea (NextGenerationEU/PRTR).

Asimismo, los autores agradecen el apoyo del Consorcio del Centro Asociado de la UNED en Ponferrada, INTECCA, así como del grupo de investigación UNED CiberCSI y del grupo de innovación UNED CiberGID.

Referencias

[1] Daniel Blanco-Aza et al. “Privacy Analysis in Mobile Apps and Social Networks using AI Techniques”. En: *Fourth Intelligent Cybersecurity Conference (ICSC 2024)*. Valencia, Spain, 2024.

[2] European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. https://eur-lex.europa.eu/eli/reg/2016/679/oj (Fecha de último acceso: 20 de mayo de 2025). 2016.

[3] P. K. Smith et al. “Cyberbullying: Its nature and impact in secondary school pupils”. En: *Journal of Child Psychology and Psychiatry* 49.4 (2008), págs. 376-385. DOI: 10.1111/j.1469-7610.2007.01846.x.

[4] Y. Feng et al. “AC-Net: Assessing the Consistency of Description and Permission in Android Apps”. En: *IEEE Access* 7 (2019), págs. 57829-57842. DOI: 10.1109/ACCESS.2019.2912210.

[5] A. Jain et al. “ATLAS: Automatically Detecting Discrepancies Between Privacy Policies and Privacy Labels”. En: *arXiv* (2023). eprint: 2306.09247. URL: http://arxiv.org/abs/2306.09247.

[6] K. Zhao et al. “Demystifying Privacy Policy of Third-Party Libraries in Mobile Apps”. En: *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. 2023, págs. 1583-1595. DOI: 10.1109/ICSE48619.2023.00137.

[7] A. J. Bhatt, C. Gupta y S. Mittal. “iABC-AL: Active learning-based privacy leaks threat detection for iOS applications”. En: *Journal of King Saud University - Computer and Information Sciences* 33.7 (2021), págs. 769-786. DOI: 10.1016/j.jksuci.2018.05.008.

[8] M. S. Rahman et al. “PermPress: Machine Learning-Based Pipeline to Evaluate Permissions in App Privacy Policies”. En: *IEEE Access* 10 (2022), págs. 89248-89269. DOI: 10.1109/ACCESS.2022.3199882.

[9] A. Xiang, W. Pei y C. Yue. “PolicyChecker: Analyzing the GDPR Completeness of Mobile Apps’ Privacy Policies”. En: *2023 ACM SIGSAC Conference on Computer and Communications Security*. 2023, págs. 3373-3387. DOI: 10.1145/3576915.3623067.

[10] B. Andow et al. *PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play*. https://www.usenix.org/conference/usenixsecurity19/presentation/andow. 2019.

[11] Y. Yang, X. Du y Z. Yang. “PRADroid: Privacy Risk Assessment for Android Applications”. En: *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)*. 2021, págs. 90-95. DOI: 10.1109/CSP51677.2021.9357608.

[12] H. N. Atapattu et al. “A Sensitive Data Leakage Detection and Privacy Policy Analyzing Application for Android Systems (PriVot)”. En: *2021 3rd International Conference on Advancements in Computing (ICAC)*. 2021, págs. 300-304. DOI: 10.1109/ICAC54203.2021.9671075.

[13] S. Li et al. “Collect Responsibly But Deliver Arbitrarily?: A Study on Cross-User Privacy Leakage in Mobile Apps”. En: *2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022, págs. 1887-1900. DOI: 10.1145/3548606.3559371.

A Review of "Analysis and effectiveness of deeper levels of SVD on performance of hybrid DWT and SVD watermarking"

Tanya Koohpayeh Araghi and David Megías
Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Spain
CYBERCAT-Center for Cybersecurity Research of Catalonia
{tkoohpayeharaghi, dmegias}@uoc.edu

***Abstract*—In this paper, we proposed a novel technique called SVD2 to enhance the imperceptibility and robustness of watermarked images. Experimental results show a remarkable increase in both imperceptibility and robustness in the discrete wavelet transform (DWT) and the second level of singular value decomposition (SVD) (proposed technique), compared to the first level of SVD (traditional technique). In addition, the achieved result shows that the proposed scheme offers the highest imperceptibility in the LL sub band (more than 60 dB on average PSNR), with satisfactory robustness against noise attacks. For the HH sub-band, strong robustness against all types of tested of attacks is obtained, though its imperceptibility is slightly lower than the achieved PSNR in the LL sub-band. In the HH sub band, an average growth of 5 dB in PSNR and 2 percent in NC can be observed from the proposed technique in comparison to the traditional technique. These results make SVD 2 a good candidate for content protection, especially for medical images.**

***Index Terms*—Hybrid Image Watermarking, DWT, SVD1, SVD2, Security, Robustness, Imperceptibility**

Tipo de contribución: Investigación ya publicada [1]

I. Introduction

Imperceptibility and robustness are two important factors in every watermarking schemes. In order to achieve maximum quality for these factors the current schemes recommended subsidiary techniques such as artificial intelligence techniques or extra transformations resulting time, cost and energy consumption. Our proposed technique can achieve even better quality in comparison to these schemes by the idea of the second level of SVD. In the paper [1] an analysis is made for the performance of both proposed schemes in all DWT sub bands in terms of robustness, imperceptibility, security and capacity and also compared them to some of the recent researches. For the robustness test, here we referred only to two HH and LL sub bands because of space limitation. Experimental results show a considerable enhancement in SVD 2 in comparison to the other schemes.

II. Methodology

In SVD decomposition, the largest singular values carry most of the energy of the image; hence, embedding a watermark into these coefficients results in a robust watermarked image. Inspired by this idea, we designed and implemented two hybrid watermarking schemes, one, in the second level of SVD, and another with only one level of SVD, to investigate the robustness and the imperceptibility of the watermarked images for both schemes. In one scheme, the host image is transformed using DWT and SVD to embed the watermark image in the singular values of the host image using all four DWT sub-bands. The other scheme uses two levels of SVD in such a way that, firstly, the host image is decomposed by DWT and each sub-band of it, is divided into non-overlapping n $\times$ n matrices, then SVD is applied on each n $\times$ n blocks to collect all largest singular values located in S (1, 1) of each block in a separate matrix in order to hide the watermark in the singular values of this matrix, after applying the second level of SVD. We denote "SVD 2" as the use of a second level of SVD, whereas "SVD 1" refers to a single level of the SVD. Our effort is to analyze the efficiency of SVD 1 and SVD 2. Therefore, these two schemes are designed and implemented based on approximate matching algorithms, such as watermark preparation and embedding, with the same resources like memory and CPU speed.

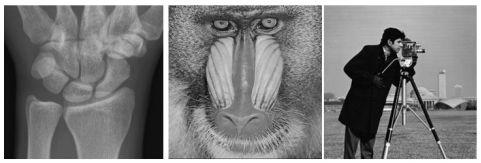


Fig. 1. Sample of tested images (from left to right): X-Ray and Baboon as host images, and cameraman as watermark.

III. Experimental results

In this section experimental results are presented for two samples of images, one medical image X-Ray of size 1024×1024 and one non-medical image Baboon of size 512×512, as shown in Figure 1. The original watermark image is Cameraman 64×64.

A. Test of Imperceptibility

In this test, the PSNR of the watermarked images for both SVD 1 and SVD 2 schemes in two medical and non-medical samples are illustrated. Table 1 shows the comparison of imperceptibility for the medical image X-ray and non medical image Baboon in the proposed SVD 1 and SVD 2 for each peer DWT sub-band.

B. Test of Security

The watermark extraction permission is only given to the authorized parties. As a result, both false positive and false negative effects can be detected and prevented. In this test, we considered a scenario in which an attacker tries to extract

TABLE I
COMPARISON OF PSNR IN SVD 1 AND SVD 2 FOR X-RAY AND BABOON.

Difference of SVDs	X-RAY			
	LL	HL	LH	HH
SVD1	53.6056	39.1522	40.3121	41.3604
SVD2	65.5593	43.5478	44.697	46.9879
Difference of SVDs	BABOON			
	LL	HL	LH	HH
SVD1	61.8291	39.3697	40.388	41.3954
SVD2	77.6624	43.5377	44.84	46.9441

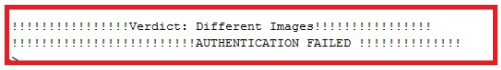


Fig. 2. Result of false positive test in both SVD 1 and SVD2.

his own watermark by injecting the fake U and V matrices. Figure 2 shows the result of our authentication system for the unauthorized user.

C. Test of payload

As mentioned before, the original size of the watermark is 64×64 but in the preprocessin phase we increased it redundantly to reach to the size of the cover image. Moreover, an eight bit digital signature is hidden to the host to prevent the false positive effect.

D. Test of robustness

In order to show the ability of the watermarked image to withstand against signal processing and geometric attacks, the watermarked image is exposed on the most recognized attacks mentioned in Tables 2 and 3, and the corresponding results are shown for both SVD1 and SVD2 schemes for each of medical and non-medical samples in their peer DWT sub-bands.

TABLE II
COMPARISON OF ROBUSTNESS (NC) BETWEEN SVD1 AND SVD2 IN LL SUB-BAND FOR MEDICAL AND NON-MEDICAL IMAGES.

Attacks	LL Sub-band, Cover: X-ray (1024 × 1024)		LL Sub-band, Cover: Baboon (512 × 512)	
	SVD1	SVD2	SVD1	SVD2
No attack	1	1	1	1
GAUSSIAN 0.01	0.8829	0.9954	0.9375	0.997
Average Filtering 3*3	0.9944	0.9967	0.8548	0.9953
CROP 1/4	-0.2806	0.6258	0.223	0.6809
Salt and Pepper 0.01	0.982	0.9976	0.9914	0.9956
SCALING 512-256-512	0.9943	0.9969	0.9338	0.9965
Histogram Equalization	0.8657	0.2521	0.321	0.2038
SPECKLE 0.01	0.9843	0.997	0.9847	0.9968
MEDIAN 3*3	0.9927	0.9966	0.9007	0.9849
Rotation 50	0.1054	0.5171	0.3269	0.6087
Gamma Correction 0.5	NAN	NAN	NAN	NAN
Compression 50	NAN	NAN	NAN	NAN

IV. COMPARISON AND DISCUSSION

In this section, the proposed SVD1 and SVD2 schemes are compared with several recent proposed counterparts. The difference of imperceptibility between the proposed schemes and these schemes for Baboon image is shown in Table 4. It is observed that the proposed SVD2 scheme is more imperceptible than all other schemes.

TABLE III
COMPARISON OF ROBUSTNESS (NC) BETWEEN SVD1 AND SVD2 IN HH SUB-BAND FOR MEDICAL AND NON-MEDICAL IMAGES.

Attacks	HH Sub-band, Cover: X-ray (1024 × 1024)		HH Sub-band, Cover: Baboon (512 × 512)	
	SVD1	SVD2	SVD1	SVD2
No attack	1	1	1	1
GAUSSIAN 0.01	0.9548	0.9855	0.9716	0.9929
Average Filtering 3*3	0.9995	0.9999	0.9753	0.9929
CROP 1/4	0.9993	0.9997	0.9697	0.9915
Salt and Pepper 0.01	0.9861	0.9928	0.9965	0.9985
SCALING 512-256-512	0.9993	0.9997	0.9699	0.9916
Histogram Equalization	0.9984	0.9995	0.9761	0.9944
SPECKLE 0.01	0.9913	0.9974	0.9965	0.999
MEDIAN 3*3	0.9995	1	0.9854	0.9956
Rotation 50	0.9995	0.9998	0.9929	0.9971
Gamma Correction 0.5	0.9992	0.9997	0.9629	0.9886
Compression 50	0.9993	0.9998	0.977	0.9934

TABLE IV
COMPARISON OF IMPERCEPTIBILITY ACCORDING TO PEAK SIGNAL TO NOISE (PSNR) AMONG SVD1, SVD2 AND OTHER RELATED WORKS FOR BABOON IMAGE.

Scheme	SVD+DWT +ABC	DWT+SVD +PSO	Our DWT+SVD1	Our DWT+SVD2
PSNR (dB)	40.0256	43.2199	41.3954	46.9441

V. CONCLUSION

A subtle design and implementation of algorithms can decrease the computational burden and cost, while at the same time; increase the performance (imperceptibility and robustness) of watermarking algorithms. Of course, depending on the application, the requirements can be varied between three specifications of robustness, capacity and imperceptibility. As we have shown in this paper, only taking a few extra steps, the implemented schemes can be more efficient than those using AI base subsidiary algorithms to achieve optimum values for imperceptibility and robustness.

ACKNOWLEDGMENT

The authors acknowledge the funding obtained from the "SECURING" project (PID2021-125962OB-C31) funded by the Ministry of Science and Innovation, the Agencia Estatal de Investigación, and the European Regional Development Fund (ERDF), and by the ARTEMISA International Chair of Cybersecurity (C057/23) and the DANGER Strategic Project of Cybersecurity (C062/23), both funded by the Spanish National Institute of Cybersecurity through the European Union – NextGenerationEU– and the Recovery, Transformation, and Resilience Plan.

REFERENCES

[1] T. K. Araghi and D. Megías, "Analysis and effectiveness of deeper levels of svd on performance of hybrid dwt and svd watermarking," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 3895–3916, 2024.

Una revisión de "Uso del aprendizaje basado en juegos y la computación cuántica para mejorar las competencias STEAM en la educación K-16"

Daniel Escanez-Exposito
Universidad de La Laguna
Tenerife, Spain
jescanez@ull.edu.es

Javier Correa-Marichal
Universidad de La Laguna
Tenerife, Spain
javier.correa.18@ull.edu.es

Pino Caballero-Gil
Universidad de La Laguna
Tenerife, Spain
pcaballe@ull.edu.es

***Resumen*—La computación cuántica es un dominio emergente y en rápida expansión que atrae a personas científicas e ingenieras. Reconociendo las limitaciones de los enfoques educativos convencionales para preparar adecuadamente a los individuos en su incursión en esta área, esta investigación introduce un novedoso juego de mesa llamado "Qubit: The Game", cuyo objetivo es doble: fomentar el entusiasmo por la computación y criptografía cuánticas y mejorar la comprensión de conceptos fundamentales dentro de esta disciplina. Este documento proporciona explicaciones sobre la razón detrás de la elección del formato de juego de mesa, su diseño y sus mecánicas, así como la metodología seguida durante su desarrollo. Además, presenta un primer análisis realizado para evaluar el impacto del juego diseñado en la percepción, el interés y el conocimiento fundamental de la computación cuántica entre estudiantes K-16. Los resultados de esta investigación demuestran de manera unívoca que el juego desarrollado es una herramienta eficaz para cultivar el disfrute y facilitar la comprensión de conocimientos esenciales en un tema tan complejo como la computación cuántica. De hecho, la efectividad de este juego también resalta su potencial para introducir a los estudiantes en diferentes temas relacionados con STEAM.**

***Index Terms*—computación cuántica, criptografía cuántica, aprendizaje basado en juegos, educación K-16.**

Tipo de contribución: *Investigación ya publicada* [1]

I. Introducción

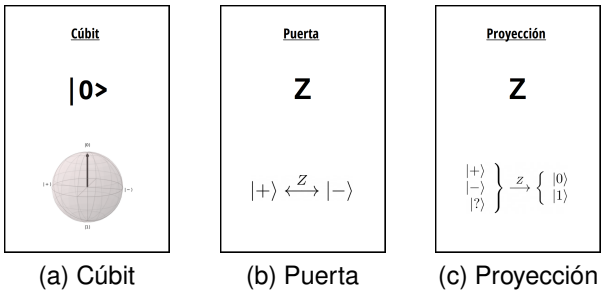
La computación cuántica representa una amenaza para la ciberseguridad, ya que puede romper los algoritmos criptográficos actuales. Para facilitar la adaptación a la era post-cuántica, se presenta "Qubit: The Game", un juego educativo que enseña conceptos clave como cúbits, puertas cuánticas y proyecciones [2]. A través de las cartas que representan estos conceptos, los jugadores construyen una clave binaria, explorando así fundamentos esenciales de las amenazas para la criptografía pre-cuántica y sus soluciones cuánticas [3].

En la computación cuántica, la unidad fundamental de información es el cúbit, que a diferencia del bit clásico, puede existir en una combinación lineal de los estados de la base computacional $|0\rangle$ y $|1\rangle$ que tienen la forma $|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle$ con $\alpha, \beta \in \mathbb{C}$ y $||\alpha||^2 + ||\beta||^2 = 1$. De esta manera, el estado cuántico $|0\rangle$ corresponde a $\alpha = 1$ y $\beta = 0$, mientras que el caso opuesto ocurre para $|1\rangle$. Este principio se conoce como superposición, permitiendo que el cúbit adopte simultáneamente ambos valores con ciertas probabilidades [4]. La representación visual de este fenómeno se logra mediante la esfera de Bloch, donde el estado del cúbit se ilustra como un vector dentro de una esfera, facilitando la comprensión de las probabilidades asociadas a cada estado al momento de la medición.

Las operaciones en computación cuántica se llevan a cabo mediante puertas cuánticas, que manipulan el estado de los cúbits [5]. Algunas de las más relevantes incluyen las puertas de Pauli (X y Z), la puerta de Hadamard (H) y la puerta $SWAP$, cada una de las cuales realiza transformaciones específicas sobre los estados cuánticos. En los algoritmos y protocolos criptográficos cuánticos, los cúbits son sometidos a estas operaciones hasta que se obtiene un resultado observable, lo cual se logra mediante la proyección del estado del cúbit sobre un eje específico, causando el colapso del estado en un valor determinado.

Otro concepto clave es el entrelazamiento cuántico, un fenómeno exclusivo de la computación cuántica en el cual dos o más cúbits comparten un estado global, de modo que la medición de uno determina instantáneamente el estado del otro, sin importar la distancia que los separa [6]. Esta propiedad es fundamental para muchos protocolos cuánticos, incluyendo la criptografía cuántica y la computación distribuida, ya que permite la correlación de resultados independientemente de la distancia [7].

Por último, se expone la medición cuántica como el proceso mediante el cual un cúbit colapsa en uno de los estados que conforman una determinada base. Esto ocurre con una probabilidad dependiente de sus coeficientes complejos, eliminando así cualquier superposición previa y obteniendo un resultado clásico, es decir, un único bit de información [8].



(a) Cúbit (b) Puerta (c) Proyección

Figura 1. Cartas utilizadas en el juego para ilustrar estos conceptos

II. Descripción General

"Qubit: The Game" es un juego de cartas en el que el objetivo del jugador consiste en ser el primero en formar una clave binaria, denominada contraseña, seleccionada aleatoriamente antes de iniciar la partida. Para lograrlo, los jugadores deben competir para colocar estratégicamente sobre la mesa una combinación de cartas de cúbit que forme la secuencia de contraseña objetivo. Para lograr esto, deberán utilizar las cartas de cúbit puerta y proyección expuestas antes. La Ec. (1) muestra el lenguaje que contiene todas las secuencias posibles de cartas jugadas en el tablero que permiten ganar la partida, donde $key \in (0|1)^*$ representa la cadena binaria objetivo.

$$L = \{w : (0|1| + |-)^* \; key \; (0|1| + |-)^*\} \tag{1}$$

El juego debe ser jugado por dos a seis participantes, siendo cuatro el número ideal para una partida estándar. La duración de cada partida dependerá de la experiencia de los jugadores con el juego. Así, las primeras partidas pueden desarrollarse de forma algo más lenta, aunque se agilizarán a medida que se adquiera un mayor conocimiento de las reglas y la mecánica. Hasta cuatro jugadores, se recomienda establecer el tamaño de la contraseña objetivo en 3 o 4, dependiendo de si se desea una partida corta (10-20 minutos) o mediana (15-30 minutos). Para partidas de cinco jugadores o más, se aconseja una contraseña objetivo de longitud 3, con el fin de lograr partidas intermedias o incluso largas (+30 minutos). Los elementos esenciales para el desarrollo del juego incluyen la baraja de cartas de "Qubit: The Game", un dado de seis caras y una moneda, para generar las mecánicas de aleatoriedad requeridas en el juego.

III. Estudio Piloto

La investigación sobre el impacto del juego propuesto se llevó a cabo con una cohorte de 96 estudiantes matriculados en el primer curso de Bachillerato y 19 estudiantes del primer curso de enseñanza secundaria de un centro educativo [9].

Con el fin de validar el cumplimiento de los distintos objetivos del juego, se realizó un estudio estadístico basado en 3 hipótesis nulas, a partir de las respuestas a la encuesta de 10 preguntas que contestaron los estudiantes. En primer lugar, en relación con la mejora de la motivación para aprender computación cuántica a través del juego, la primera hipótesis nula planteada en el estudio de los datos obtenidos de las primeras cuatro preguntas de la encuesta se denota como H1 a continuación. Mediante las preguntas Q5 a Q7, la segunda hipótesis nula relacionada con el juego se denomina H2. Por último, la evaluación de la mejora en la comprensión de conceptos fundamentales de computación cuántica se realiza con las preguntas Q9 y Q10, lo que permite probar la hipótesis nula H3:

- **H1** No existe diferencia significativa en la actitud de los estudiantes hacia la computación cuántica antes y después de la exposición al juego.
- **H2** No existe diferencia significativa en el interés de los estudiantes por jugar al juego antes y después de la exposición al mismo.
- **H3** No existe diferencia significativa en los conocimientos básicos de computación cuántica de los estudiantes antes y después de la exposición al juego.

Se han utilizado pruebas t de Student para dos muestras pareadas en esta investigación, puesto que todas las personas respondieron la encuesta dos veces. Los resultados de las pruebas realizadas en las preguntas Q1 a Q4 respaldan el rechazo de la hipótesis nula H1, concluyendo que existe una diferencia significativa en la actitud de los estudiantes hacia la computación cuántica antes y después de jugar. De igual forma, los resultados de las preguntas Q5, Q6 y Q7 avalaron el rechazo de la segunda hipótesis nula H2, determinando que existe una diferencia significativa en el interés de los estudiantes por jugar antes y después de probar el juego. En lo referente a la pregunta Q8, se observa que, tras jugar, se mantiene el interés inicial en los conceptos de cúbit, puertas cuánticas y entrelazamiento, mostrando menor interés en el concepto de esfera de Bloch. Finalmente, las preguntas Q9 y Q10 confirman la eficacia del juego para adquirir conocimientos relacionados con conceptos básicos de computación cuántica.

IV. Conclusiones

Este juego de mesa, llamado "Qubit: The Game", busca despertar el interés por la computación cuántica y reforzar conceptos clave como cúbit, superposición, esfera de Bloch, puertas cuánticas, proyección, entrelazamiento y decoherencia. Un estudio realizado con estudiantes de enseñanza secundaria mostró que la dinámica del juego resulta efectiva para motivar y mejorar la comprensión de estos temas, siendo apto para quienes no están familiarizados con la disciplina.

Se proponen dos extensiones: una incorpora el uso del eje Y (añadiendo nuevos estados de cúbit y proyecciones) y la otra introduce elementos de criptografía mediante la aplicación de algoritmos de Distribución Cuántica de Claves (QKD). Estas modificaciones, planteadas como expansiones, permiten que los jugadores exploren estrategias más complejas sin alterar la curva de aprendizaje básica. Además, se contempla su lanzamiento como iniciativa de código abierto para docentes, estudiantes y aficionados a los juegos de cartas.

Referencias

[1] Escanez-Exposito, D., Correa-Marichal, J., Caballero-Gil, P., Using Game-Based Learning and Quantum Computing to Enhance STEAM Competencies in K-16 Education. IEEE Transactions on Education, vol. 67, no. 6, pp. 807-816, Dec. 2024.

[2] Singh, C., Helping students learn quantum mechanics for quantum computing. In: AIP Conference Proceedings, vol. 883(1), pp. 42–45. American Institute of Physics (2007).

[3] Escanez-Exposito, D., Caballero-Gil, P., Martín-Fernández, F., Interactive simulation of Quantum Key Distribution protocols and application in Wi-Fi networks. Wireless Networks, Springer (2023).

[4] Seskir, Z.C. et al. Quantum games and interactive tools for quantum technologies outreach and education. Optical Engineering, 61(8), pp. 081809-081809 (2022).

[5] McMahon, D.: Quantum computing explained. John Wiley & Sons (2007).

[6] He, Y., Zha, S., He, W., A literature review of quantum education in K-12 level. Innovate Learning Summit, pp. 418-422 (2021).

[7] Satanassi, S., Ercolessi, E., Levrini, O., Designing and implementing materials on quantum computing for secondary school students: The case of teleportation. Physical Review Physics Education Research, 18(1), 010122 (2022).

[8] Singh, C.: Interactive learning tutorials on quantum mechanics. American Journal of Physics, 76(4), pp. 400-405 (2008).

[9] Foti, C., Anttila, D., Maniscalco, S., Chiofalo, M.L., Quantum physics literacy aimed at K12 and the general public. Universe, 7(4), p. 86 (2021).

Securización frente a decompilaciones: enfoque en aplicaciones móviles Android de Machine Learning

Luis Manuel Muñoz Pérez
EUPT, Universidad de Zaragoza
Teruel, España
l.munoz@unizar.es

Raúl Igual Catalán
EUPT, Universidad de Zaragoza
Teruel, España
rigual@unizar.es

Resumen- **Este estudio presenta un método de cifrado simétrico para ser utilizado en aplicaciones móviles Android que operen modelos de Machine Learning de manera local. A través de un proceso basado en encriptación inicial, ofuscación de la clave y desencriptación final, se consigue mejorar la confidencialidad, integridad y disponibilidad de los archivos presentes en este tipo de aplicaciones móviles. El método de cifrado simétrico propuesto ha sido probado en 10 aplicaciones móviles. En este estudio se analizan los resultados de una de ellas durante varios años. Se puede determinar que la técnica presentada mejora la seguridad, manteniendo la fiabilidad de las aplicaciones móviles y siendo robusta frente a decompilaciones de código que pretendan extraer la clave.**

***Index Terms*- cifrado simétrico, decompilación, aplicaciones móviles, Machine Learning**

Tipo de contribución: *Investigación original*

I. Introducción

Con el incipiente uso de la Inteligencia Artificial en todos los aspectos de la sociedad, tanto en el ámbito industrial como doméstico, cobra mayor importancia el problema de la seguridad. Es necesario mantener la confidencialidad e integridad de las implementaciones que utilizan Inteligencia Artificial y están disponibles para el público general. Entre ellas, se encuentran los modelos de Machine Learning integrados de forma nativa en aplicaciones móviles, que han de operar sin conexión a internet, de forma integrada y con unos recursos limitados.

La opción más utilizada actualmente en cuanto a uso de modelos de Machine Learning en aplicaciones móviles Android, es el almacenamiento del modelo en el servidor por parte del proporcionador de servicios [1]. De este modo, para utilizar esos modelos, se realiza un proceso de carga y descarga de datos al servidor. En muchos casos, esto supone el envío y recepción de datos sensibles. Así mismo, se emplea una cantidad elevada de recursos, resultando en una disponibilidad inferior a la que se obtendría si todo esto se realizara de manera local, sin comunicación con el servidor.

La ejecución de los modelos de Inteligencia Artificial en el propio dispositivo del usuario economiza recursos, evita la compartición de la información y permite una mayor disponibilidad al realizar todas las operaciones en un entorno local. Por esta razón, resulta de especial interés contar con una metodología para ayudar a la mejora de la seguridad en aquellas implementaciones que eviten la transmisión de datos, manteniendo la confidencialidad e integridad del modelo utilizado.

Este mayor nivel de confidencialidad puede conseguirse mediante la encriptación de los ficheros que contienen la información sensible. De este modo, cuando la aplicación sea decompilada, los ficheros se encontrarán encriptados y, salvo que se consiga desencriptarlos, la información guardada permanecerá oculta para los posibles intrusos. Para ello, será de especial importancia la salvaguarda de la clave de desencriptado, que será ofuscada y embebida en la compilación de la aplicación para evitar su posterior obtención [2].

Este artículo presenta un método para la securización de la información presente en local en las aplicaciones móviles, estando especialmente dirigido a aquellas que ejecutan modelos de Machine Learning. La técnica incluída en este artículo es de aplicación para la securización de cualquier tipo de fichero que vaya a ser integrado en aplicaciones móviles, aunque se centra especialmente en aplicaciones basadas en el sistema operativo Android, por ser el más utilizado. Fuera del alcance de este estudio quedan las aplicaciones implementas para el sistema iOS.

II. Trabajos Relacionados

Existen muy pocos trabajos previos publicados en este ámbito. El trabajo de Sun et al. [3] presenta una descripción detallada sobre el problema de la securización de aplicaciones móviles con Machine Learning, sin presentar aplicaciones prácticas concretas. Por su parte, otros autores han propuesto algunos métodos publicados en repositorios de libre acceso [4,5]. Sin embargo, a diferencia de estos trabajos, el método propuesto en este artículo no hace uso de servidores externos y presenta una solución completa para la securización frente a decompilaciones.

III. Metodologías de encriptadoo

El encriptado a utilizar deberá garantizar una adecuada seguridad frente al desencriptado. Las principales metodologías empleadas son las siguientes:

• Cifrado simétrico: se usa la misma clave para el encriptado y desencriptado.

• Cifrado asimétrico: existen dos claves, una pública y una privada (que debe quedar oculta). La obtención de la clave privada necesitaría de un complejo cálculo matemático para su obtención sobre la clave pública. Los cálculos suelen estar basados en exponenciación modular, pruebas de primalidad o criptografía de curva elíptica.

Por un lado, es posible optar por la implementación de un encriptado asimétrico. Esto, en primer lugar, evitaría ofuscar la clave pública como ocurre con el cifrado simétrico.

En el caso de utilizar un cifrado simétrico existe el problema que supone la salvaguarda de la clave empleada [6]. Si bien existe una inseguridad asociada al almacenamiento de

la clave en la aplicación, la compilación de la aplicación móvil antes de su publicación como ejecutable ofrece una posible capa de seguridad. Esto puede realizarse mediante un guardado de la clave entre las variables de compilado de la aplicación (local.properties; SDK Location). De este modo, una vez se compila la aplicación, queda ofuscada entre la estructura interna y codificada de la aplicación. Esta opción no supone una capa de seguridad elevada, dado que los contenidos de la aplicación no están lo suficientemente ofuscados de cara a una decompilación selectiva. Por este motivo, se plantea una medida adicional para proteger las claves: definirlas directamente en los archivos de variables utilizados durante la compilación de la aplicación [2]. De esta manera, se dificulta significativamente su acceso o extracción una vez que la aplicación ha sido compilada.

Este procedimiento hace que la encriptación simétrica disponga de algunas características de seguridad vinculadas a la encriptación asimétrica. Esto es, dada la ofuscación adicional de la clave, se puede ofrecer la aplicación con la clave precompilada a los usuarios que necesiten desencriptar los ficheros relacionados con la aplicación. De esta manera, el conjunto de la aplicación con la clave no pública precompilada actúa en la práctica como una clave pública disponible para cualquiera que quiera descargar la aplicación.

La metodología de ofuscación de clave simétrica puede suponer una importante mejora en la Disponibilidad de la información, al reducir considerablemente el tiempo de procesado y consumo de recursos. Esto es fundamental en ciertas aplicaciones como pueden ser la identificación en tiempo real utilizando modelos de Machine Learning.

No obstante, una decompilación avanzada podría exponer la clave ofuscada a pesar de haber aplicado la técnica expuesta. Aun así, las prácticas de compilación y ofuscación presentadas deberían mejorar considerablemente la seguridad de la clave, permitiendo un procesado rápido de los modelos encriptados, con un nivel adecuado de seguridad [2].

De este modo, el método propuesto de cifrado simétrico es útil en aplicaciones móviles basadas en modelos de Machine Learning u otras técnicas de Inteligencia Artificial en tiempo real o de operación rápida. La mayor parte de implementaciones actuales basadas en modelos almacenados en servidor hacen uso de metodologías de encriptación asimétrica. El cifrado asimétrico tiene seguridad probada y es conveniente para usar en el paradigma de datos compartidos por medio de servidor de almacenamiento, generalmente almacenamiento en la nube escalable. Por el contrario, en este estudio se opta por utilizar encriptado simétrico con ofuscación de cara al uso de modelos de Machine Learning almacenados localmente, permitiendo rapidez y bajo consumo de recursos.

Este estudio propone el uso de aplicaciones de cifrado simétrico estandarizadas y sencillas, que garantizen una rápida implementación y a su vez provean de una mejora en el nivel de seguridad de la aplicación.

Los métodos de encriptado simétrico más habituales por su eficacia, popularidad y fácil implementación son AES (Advanced Encryption Standard), 3DES y ChaCha20.

Por un lado, el método DES ha sido muy popular pero también es relativamente fácil de desencriptar [6]. Es por ello, que ha evolucionado hacia el 3DES, consistente en 3 *wrappings* del método DES; si bien, este es igualmente descrifrable, solo que más costoso computacionalmente [7].

Por otro lado, el método ChaCha20 se encuentra entre los más populares, efectivos, y rápidos. No obstante, el método AES tiene unas características de velocidad y seguridad muy similares a ChaCha20 [8]. Siendo a grandes rasgos ambas metodologías similares, y gozando de una gran popularidad, resulta conveniente hacer uso de la metodología AES para facilitar la implementación práctica de forma similar a las técnicas más habitualmente utilizadas.

IV. Etapas del método de cifrado simétrico



Fig. 1. Esquema de implementación del método de cifrado simétrico.

Los diferentes pasos del método basado en cifrado simétrico propuesto en este estudio se muestran en la Figura 1.

En primer lugar, se debe contar con el fichero encriptado a utilizar en la aplicación. Para ello, se extrae la información contenida en el fichero original, se encripta con la clave final, y se guarda en un nuevo archivo, que será el que contenga la aplicación (Sección V). Este archivo, tras ser decompilado, no debería poder descifrarse sin contar con la clave.

Por otro lado, habrá que contar con una aplicación móvil que será la que contenga el fichero generado en el paso anterior. Este método está especialmente pensado para aplicaciones que contengan modelos de Machine Learning, para permitir así su ejecución en local con seguridad.

Con la aplicación móvil desarrollada, ya se puede incluir el archivo encriptado en los recursos de la aplicación.

Se realiza la configuración para el ofuscado de la clave (Sección VI). Para ello, se utilizan principalmente las ventajas de la compilación y almacenamiento de variables fuera del código fuente, mejorando la seguridad frente a una decompilación intencionada [2].

A su vez se crea el código que realiza el desencriptado y hace uso de la clave (Sección VII).

V. Encriptación inicial de los ficheros

El encriptado y el desencriptado se realizan en aplicaciones diferentes. Esta sección aborda el proceso de encriptado, que es el primer paso del método.

Para encriptar inicialmente los ficheros, se pueden seguir diferentes aproximaciones. Se puede realizar el primer encriptado con una implementación de la técnica [9] (AES, como se ha indicado en la Sección III), por cualquier medio que lo permita, utilizando un programa en Python, Rust, C, etc.

Dado que el desencriptado (Secciones VI y VII) se desarrolla estrictamente en el entorno de programación Android (el de las aplicaciones móviles), por facilidad, en

este trabajo, se realiza también el encriptado en este entorno. Para ello, se puede usar Java o Kotlin (una aproximación más moderna sustitutiva de Java). En este trabajo la implementación se ha hecho en Kotlin.

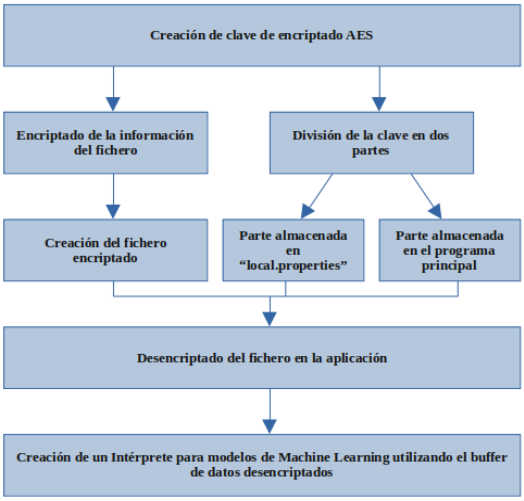


Fig. 2. Diagrama de la implementación realizada con Kotlin.

Para llevar a cabo el proceso de encriptación, en primer lugar, establecemos una clave, siendo los tamaños habitualmente utilizados 128, 192 y 256 bits, dependiendo de las necesidades de seguridad y rapidez de procesado [10]. En los ejemplos que se muestran en este trabajo se usa un tamaño de clave de 128 bits.

La clave se separa en 2 partes (Fig. 2, código en [11]):

• Una parte deberá estar contenida en la implementación de desencriptado en las variables *Gradle* de inyección de configuraciones (Sección VI) para mayor seguridad a la hora de ocultar la clave.

• La otra parte de la clave estará en el programa principal de desencriptado (Sección VII), ofuscada solamente con la configuración de seguridad utilizada por defecto en la compilación de Android Studio.

La clave se transforma en una cadena de Bytes para poderla aplicar con el formato específico de encriptación (en este caso AES).

Por otro lado, se toma el fichero a encriptar y se extrae la información contenida (Fig. 2, código en [11]). Esta información se guarda en una cadena de Bytes, para poder operar con las librerías de cifrado AES.

A partir de la información ya encriptada se genera el nuevo fichero que será almacenado en la aplicación a distribuir libremente entre toda la base de usuarios (Fig. 2, código en [11]).

Para mayor comodidad, la aplicación de encriptación, una vez ha encriptado el fichero, muestra un *prompt* para elegir como guardarlo o compartirlo. Esto no supone un riesgo importante de seguridad dado que ya se encuentra encriptado.

VI. Ofuscación y embebido de la clave

En lo que se refiere a la clave (separada en dos partes; Sección V), este estudio propone su securización, guardando parte de esta clave en el entorno de desarrollo local del desarrollador, concretamente en "local.properties". Este archivo forma parte de los *Gradle Scripts* (configuración de compilación) [12]. Este archivo de configuraciones se utiliza principalmente para proporcionar la dirección donde se encuentran las herramientas de desarrollo SDK. Así mismo, es útil para el almacenamiento de claves, al ser un archivo volátil cuyo contenido se compila antes que el código fuente. Los valores contenidos en "local.properties" se leen por el configurador o gestor de compilación (*Gradle*). Esto significa que esta información nunca está incluida en el código fuente de la aplicación, pero sí oculta en el ejecutable final [2][13].

El método también ha considerado el caso de que la implementación de las aplicaciones móviles se realice con un control de versiones como Git. En este caso, las variables contenidas en "local.properties" no se procesarían, con lo que el método seguiría siendo seguro. Tampoco se mantendrían guardadas en el caso de exportar el proyecto.

Por otro lado, la otra parte de la clave contenida en el código fuente (Sección V) seguirá estando protegida por la ofuscación nativa que se aplica en la compilación [2]. Esto se hace configurando en el archivo "build.gradle" el parámetro "minifyEnabled" como "true" [14]. En relación a esta ofuscación nativa, también se pueden editar las reglas contenidas en el fichero "proguard-rules.pro", que por defecto se encuentra vacío (toma la configuración estándar de Google) [14]. Por otro lado, los ficheros e imágenes contenidos en la aplicación son recursos y no son ofuscados nunca. Por ello, es conveniente la encriptación previa de los ficheros con información sensible que hayan de estar contenidos en la aplicación.

El modo de componer la clave final (ambas partes) a ofuscar puede verse más claramente en los *snippets* asociados a este artículo [11] (Fig. 2), donde se muestra en qué secciones de la aplicación se encuentra definida cada una de las partes.

Aun así, al estar contenida la clave en el ejecutable final de la aplicación, podría ser obtenida con ingeniería inversa empleando el tiempo y los recursos necesarios. No obstante, con la ofuscación realizada, la cantidad de tiempo y recursos a emplear aumenta considerablemente.

VII. Desencriptado final de los ficheros

El modelo encriptado se puede incluir de forma nativa en las aplicaciones móviles donde se pretenda aplicar esta metodología de seguridad. También se incluye el código necesario para su desencriptado (en los ejemplos de este estudio, utilizando las librerías de codificación de AES), como se muestra en la Fig. 2 (código en [11]).

Con la cadena de datos desencriptados, estos ya pueden ser utilizados directamente en la implementación correspondiente.

Esta metodología es especialmente útil en aplicaciones móviles que usan modelos de Machine Learning que son más proclives a trabajar offline. Para ejemplificar el modo de proceder con la información desencriptada en este tipo de aplicación concreta, la Figura 2 presenta la creación de un intérprete para modelos de Machine Learning utilizando los datos desencriptados (código en [11]). En concreto, se muestra la creación de un intérprete para modelos TensorFlow (modelos específicos a utilizar en este ejemplo).

VIII. Métodos

Para validar el método de cifrado simétrico propuesto (Secciones V, VI y VII), se ha implementado sobre 10 aplicaciones móviles realizadas en Android que utilizan técnicas de Machine Learning. Las aplicaciones son las siguientes: “Bird Seeker Antarctica”, “Bird Seeker España”, “Bird Seeker Maldives”, “Blemish Types, Skin Cancer ID”, “CloudPic Offline Meteo Predict”, “Countrynator, pic-hunt Nations”, “Eye Color Guesser”, “FotoCheck Hortalizas”, “Santa Tracker & Gift Guesser” y “WaterPic Uncover Water Secrets” [15].

De todas ellas, se ha seleccionado la aplicación “Blemish Types, Skin Cancer ID” (Figura 3) ya que es la única que cuenta con una base de usuarios suficiente para poder analizar los resultados [16].



Fig. 3. Algunos pantallazos de la aplicación móvil Android “Blemish Types, Skin Cancer ID”. Autoría original.

Para esta aplicación que incorpora el método de cifrado simétrico propuesto se han obtenido las siguientes métricas de su funcionamiento:

- Número de errores totales registrados por la aplicación antes y después de incorporar el método de cifrado simétrico.
- Número de errores por día antes y después de incorporar el método de cifrado simétrico.
- Número de errores por día y usuario antes y después de incorporar el método de cifrado simétrico.

Los errores registrados se deben a *crash* totales en la aplicación.

Así mismo, también se ha realizado una prueba de seguridad. Para ello se ha utilizado el software de decompilación Ghidra. En concreto se ha tratado de obtener la clave de descifrado decompilando la aplicación con este software. Esto ha permitido valorar si el método de cifrado simétrico es lo suficientemente robusto frente a algunas decompilaciones básicas [17].

IX. Resultados

El método de cifrado simétrico propuesto ha sido incorporado a 10 aplicaciones móviles, no resgistrándose incidencias en ninguna de ellas. La aplicación seleccionada para realizar un análisis más exhaustivo, “Blemish Types, Skin Cancer ID”, cuenta con un total de 238 usuarios activos.

La aplicación “Blemish Types, Skin Cancer ID” fue publicada el 15 de marzo de 2023 (versión actualizada). La implementación del método de cifrado simétrico propuesto en este estudio se incorporó a la aplicación el 11 de julio de 2023. La Figura 4 muestra el número de errores registrados por la aplicación desde la fecha de publicación ahora (2025).

Tabla I
ERRORES PRODUCIDOS EN LA APLICACIÓN ANTES DE LA IMPLEMENTACIÓN DE SEGURIDAD

Conjunto de Días	Usuarios	Errores/Usuario
23/03/2023	26	0.0385
27/04/2023	46	0.0217
16/05/2023	61	0.0164
23/05/2023 - 25/05/2023	71	0.0282
05/06/2023 - 10/06/2023	86	0.0349
30/06/2023 - 06/07/2023	87	0.0460
Cantidad de Errores por Usuario y Día		**0.00157**

Tabla II
ERRORES PRODUCIDOS EN LA APLICACIÓN DESPUÉS DE LA IMPLEMENTACIÓN DE SEGURIDAD

Conjunto de Días	Usuarios	Errores/Usuario
16/09/2023	89	0.0112
30/09/2023 - 12/10/2023	108	0.0370
30/10/2023 - 03/11/2023	106	0.0377
03/12/2023	112	0.0089
16/04/2024 - 20/04/2024	122	0.0246
30/04/2024 - 06/05/2024	120	0.0167
03/06/2024	121	0.0083
10/07/2024	115	0.0087
09/11/2024 - 13/11/2024	124	0.0161
15/12/2024 - 21/12/2024	130	0.0231
15/01/2025	134	0.0075
Cantidad de Errores por Usuario y Día		**0.000325**

Puede apreciarse que antes de implementar el método de cifrado simétrico, se registraron 12 errores (0.102 errores por día). Tras la implementación del cifrado simétrico, el número de errores fue de 23 (0.038 errores por día).

Las Tablas I y II presentan el número de errores por usuario activo antes y después de incorporar el método de cifrado simétrico a la aplicación, respectivamente.

A raíz de los resultados reflejados en las Tablas I y II, se puede apreciar como el número de errores por usuario activo y día disminuye al incorporar el método de cifrado simétrico propuesto. La cantidad de errores por usuario y día obtenidos como media es de 0.00157 y de 0.000325 antes y después de incorporar este método, respectivamente. Por tanto, se aprecia una disminución de 1.25 errores por cada mil usuarios y día.

El método implementado muestra una buena efectividad. Esto se logra sin alterar el funcionamiento normal de las aplicaciones con modelos de Machine Learning.

Igualmente, también se ha decompilado la aplicación con el software de decompilación Ghidra. Se ha analizado el resultado de esta decompilación y se ha observado la dificultad de extraer la clave con este [18]. Por tanto, el resultado de esta decompilación sencilla es también positivo.

Finalmente, también se ha medido el gasto de recursos de este método. La aplicación con encriptado presenta un 26% de pico de uso de CPU y de media un 18% de uso durante 1.5s. En la aplicación sin encriptado el valor de pico es del 23% y el valor medio es del 15% durante 1.3s. El procesador usado es un Mediatek Helio G99 8core Max 2.2 GHz.

X. Conclusiones

En este estudio se ha propuesto un método de cifrado simétrico para incrementar la seguridad de aplicaciones

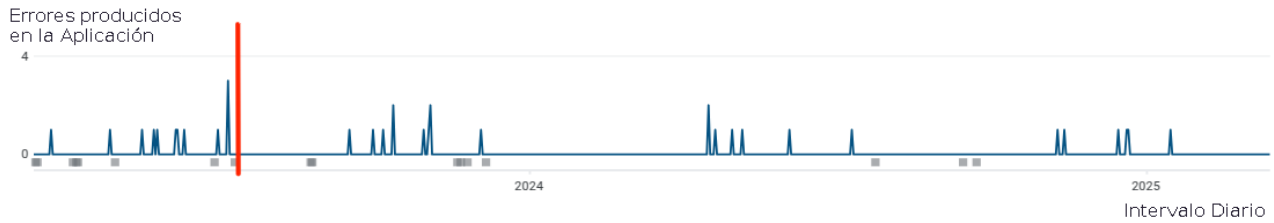


Fig. 4. Número de errores producidos durante el uso habitual de la aplicación entre usuarios activos antes y después de la implementación (separado por una línea vertical roja). Datos extraídos de Google Play Console.

móviles Android que utilizan modelos de Machine Learning en el propio dispositivo de manera local. Esta implementación es especialmente interesante para aplicaciones que deban utilizar una clave simétrica almacenada localmente, ya sea para facilitar la operatividad, por no disponer de conexión a internet, por falta de servidores de almacenamiento de datos, o por otras razones.

Este método ha sido probado en 10 aplicaciones Android distintas obteniendo resultados positivos. Se ha constatado una mejora de protección de la confidencialidad e integridad de estas aplicaciones, cuyos modelos de Machine Learning se encuentran ahora protegidos frente a una decompilación sencilla, manteniendo el nivel de disponibilidad que tenían antes de la implementación.

Se ha podido constatar que las aplicaciones no empeoraban su desempeño al incorporar el método de cifrado simétrico propuesto. Así mismo, el método propuesto parece robusto ya que no se ha conseguido acceso a la clave tras someterlo a un proceso de decompilación.

Este estudio también presenta varias limitaciones. Se desconoce si el método propuesto seguirá siendo igual de robusto cuando sea sometido a decompilaciones más avanzadas que la probada en este estudio [19]. Así mismo, los teléfonos móviles también han aumentado sus prestaciones con el paso del tiempo [20]. Esto podría haber influido en la disminución del número de errores en los años más recientes. Además, los errores registrados pueden deberse a múltiples factores, que deberían ser analizados en mayor profundidad. Sería también deseable analizar los efectos de esta metodología en más aplicaciones móviles.

Por último, es relevante destacar que, en aquellos casos en los que no se necesite ahorro de recursos, mejora del tiempo de procesamiento o funcionamiento sin comunicación con un servidor de almacenamiento de datos, la metodología propuesta no sería tan recomendable. En esos casos, las implementaciones habituales que las grandes compañías tecnológicas usan con frecuencia serían las apropiadas. Estas soluciones cuentan con infraestructura de almacenamiento en la nube gestionada por un equipo de IT, o implementaciones de servidor más simples. No obstante, la primera opción conlleva unos costes altos de mantenimiento y la segunda tiene limitaciones relacionadas con la ausencia de un equipo de IT dedicado a su gestión, entre otras.

Así, la técnica de cifrado simétrico propuesta proporciona una solución sencilla que permite mejorar la seguridad en las aplicaciones móviles que incorporen modelos de Machine Learning en el entorno local del dispositivo.

Agradecimientos

Esta iniciativa está parcialmente finanicada por el “Departamento de Empleo, Ciencia y Universidades, del Gobierno de Aragón (T49_23R)”; y parcialmente financiada por la Unión Europea a través del “Plan de Recuperación, Transformación y Resiliencia (*Next Generation*), proyecto estratégico C077/23 del Instituto Nacional de Ciberseguridad (INCIBE)”.

Referencias

[1] S. S. Qureshi, T. Ahmad, K. Rafique and Shuja-ul-islam, "Mobile cloud computing as future for mobile applications - Implementation methods and challenging issues," *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, Beijing, China, 2011, pp. 467-471, doi: 10.1109/CCIS.2011.6045111.

[2] Google, “Android Enterpirse Security Paper,” 2024.

[3] Z. Sun, R. Sun, L. Lu, and A. Mislove, “Mind Your Weight(s): A Large-scale Study on Insufficient Machine Learning Model Protection in Mobile Apps,” *30th USENIX Security Symposium*

[4] A. Verma, “Encryption and Real Time Decryption for protecting Machine Learning models in Android Applications,” arXiv preprint arXiv:2109.02270

[5] Y. Li, X. Dang, H. Tian, T. Sun, Z. Wang, L. Ma, J. Klein, and T. F. Bissyandé, “AI-driven Mobile Apps: an Explorative Study,” arXiv preprint arXiv:2212.01635

[6] J. Xu, L. Zhang, D. Lin and Y. Mao, "Recommendable Schemes of Anti-decompilation for Android Applications," *2015 Ninth International Conference on Frontier of Computer Science and Technology*, Dalian, China, 2015, pp. 184-190, doi: 10.1109/FCST.2015.76.

[7] S. Radhakrishnan and C. M. Travieso-Gonzalez, *Biometrics and Cryptography*, London, UK: IntechOpen, 2024.

[8] C. Paar, J. Pelzl and T. Güneysu, *Understanding Cryptography: From Established Symmetric and Asymmetric Ciphers to Post-Quantum Algorithms*, 2nd ed. Cham, Switzerland: Springer Nature, 2024.

[9]Cryptography. https://developer.android.com/privacy-and-security/cryptography (accedido el 17/02/2025).

[10] R. S. Danturthi, *Database and Application Security: A Practitioner's Guide*, Boston, MA: Addison-Wesley Professional, 2024.

[11] SnS. https://gitlab.com/lmmunoz/sn/-/snippets (accedido el 10/05/2025).

[12]Writing Build Scripts. https://docs.gradle.org/current/userguide/writing_build_scripts.html (accedido el 17/02/2025).

[13] U. Kargén, N. Mauthe, and N. Shahmehri, "Android decompiler performance on benign and malicious apps: an empirical study," *Empirical Software Engineering*, vol. 28, no. 2, p. 48, Feb. 2023. doi: 10.1007/s10664-022-10281-9.

[14] Shrink, obfuscate, and optimize your app. https://developer.android.com/build/shrink-code (accedido 17/02/2025).

[15] Apps Summary. https://www.appbrain.com/dev/Turion+Development (accedido el 17/02/2025).

[16]Blemish Types, Skin.Cancer ID. https://play.google.com/store/apps/details?id=com.FotoCheckSkin (accedido el 17/02/2025).

[17] N. Mauthe, U. Kargén, and N. Shahmehri, "A Large-Scale Empirical Study of Android App Decompilation," *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2021. doi: 10.1109/SANER50967.2021.00044.

[18] Y. Cao, R. Zhang, R. Liang, and K. Chen, "Evaluating the Effectiveness of Decompilers," in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2024)*, 2024

[19] T. Sutter, T. Kehrer, M. Rennhard, B. Tellenbach and J. Klein, "Dynamic Security Analysis on Android: A Systematic Literature Review," in *IEEE Access*, vol. 12, pp. 57261-57287, 2024, doi: 10.1109/ACCESS.2024.3390612.

[20] M. P. Singh and M. K. Jain, "Evolution of Processor Architecture in Mobile Phones," *International Journal of Computer Applications*, vol. 90, no. 4, pp. 34-39, Mar. 2014, doi: 10.5120/15564-4339

A Review of RBD24 : A Labelled Dataset with Risk Activities Using Log Application Data

Santiago Escuder Folch, Albert Calvo, Nil Ortiz, Josep Escrig, Maxime Compastié
i2CAT Foundation
Carrer del Gran Capità, 2, 4
Barcelona, Spain
{santiago.escuder, albert.calvo, nil.ortiz, josep.escrig, maxime.compastie}@i2cat.net

***Abstract*—RBD24 dataset is an open source resource for analysing risk activities in cybersecurity. Collected over two weeks at a Spanish university, it focuses on early-stage attack behaviours through multimodal application logs. Designed for User and Entity Behaviour Analytics (UEBA) and risk assessment, RBD24 provides ground-truth data on threats like cryptocurrency activities, outdated software, P2P file sharing, and phishing. These risks are identified via intrusion detection systems (IDS) and experimental campaigns, with indicators extracted from DNS, HTTP, SSL, and SMTP logs.**

***Index Terms*—Cybersecurity, Risk Activities, Risk Behaviour, Application Logs**

Contribution: *Research already published*

I. Introduction

This extended abstract summarises our publication RBD24: A Labelled Dataset with Risk Activities Using Log Application Data published in Elsevier Computer & Security Risk journal, in which the dataset Risk Activities Dataset 2024 (RBD24) is presented [1]. The development of data-driven approaches in cybersecurity is a significant challenge, requiring high quality datasets for training and benchmarking models. In this context, practitioners often rely on open source datasets due to the difficulty of obtaining proprietary ones. However, the number of publicly available datasets for benchmarking data-driven approaches is extremely limited, as the data is sensitive and subject to stringent privacy and security regulations. The ever-evolving nature of the threat landscape in cybersecurity requires novel datasets that capture emerging and previously untracked threats. Unfortunately, many works in threat detection still rely on outdated datasets due to data scarcity. As threats evolve, so too must the data collected by monitoring tools. Relying on older datasets not only risks omitting new patterns but also reduces the realism of experiments. Simulated environments used to generate data often lack the complexity and unpredictability of real-world scenarios, further limiting the applicability of such datasets. Moreover, most publicly available datasets over-represent network traces, particularly those derived from PCAP files and network traffic, primarily collected through NAC and NIDS-based monitoring such as Aposemat IoT-23 [2]. Nevertheless, in recent years, the cybersecurity domain has evolved toward more holistic approaches, utilising SIEM platforms to gain better insights by integrating diverse data sources, including application logs, network traces, and endpoint data. This overemphasis neglects the diversity of data sources, especially those found in application logs. The introduction of RBD24 addresses this gap by focusing on multimodal application logs, capturing a broader range of data modalities and offering a more realistic reflection of diverse cybersecurity threats. In addition to its contribution to User and Entity behaviour Analytics (UEBA) for threat detection, the dataset supports applications in risk behaviour assessment, paving the way for more effective detection techniques. To bridge the gap in UEBA analysis using ML-driven techniques, a novel open source dataset called the RBD24 has been created. This dataset comprises a collection of observed risk activities from real users at a Spanish university over a two-week period, where different multimodal application logs are preprocessed to build samples that accurately represents user's behaviour. To generate the dataset, a set of sensors were deployed across the enterprise network, providing visibility into the activities of nearly two thousand different users. The key contributions of our proposed work are as follows:

- **Catalog of Labelled Risk Activities.** The dataset contains a set of labelled risk activities extracted from real users, providing a novel ground truth dataset to support UEBA approaches.
- **User-centric multimodal Application Logs.** The RBD24 dataset is built using application logs that represent the highest abstraction level in the OSI model. Specifically, it includes information from DNS, HTTP, SSL, and SMTP protocols. This multimodality allows us to profile different entities from various perspectives.
- **Proposed Set of Indicators.** The proposed dataset preprocesses various application logs from users and constructs a set of 83 distinct indicators. These indicators can be divided in Temporal, Network Usage and Communication.

The dataset is available in the following Zenodo repository. [1]

II. The Dataset Generation Architecture

This section outlines the proposed system architecture for generating various risk activity datasets using application logs. Specifically, the application logs contain information regarding events that occur in the entity applications, categorised as Layer 7 of the OSI Model. By analysing these application logs, it is possible to correlate concrete user-level activities. For example, web browsing activities can be determined by analysing HTTP and DNS logs, while email-related activities, such as sending or receiving emails, can be inferred from SMTP application log data. To associate these application logs

[1]DOI RBD24 dataset: https://doi.org/10.5281/zenodo.13787590

with risk activities, two different heuristics are proposed to build the ground truth. The first heuristic involves identifying risk activities by using an on-premises IDS to determine logs that contain Indicators of Compromise (IOC) associated with specific activities. These activities are: *Cryptocurrency*, *Outdated software*, *P2P* and *Non-encrypted* . The second heuristic involves the generation of phishing emails and analysing the users who interact with these artifacts. To collect and preprocess the application logs, a dataset generation pipeline has been implemented. This pipeline collects logs from the source, preprocesses logs from different applications, and refines the resulting datasets.

A. Dataset Structure and Proposed Indicators

Each risk activity in the RBD24 dataset is represented by a set of samples, which are organised according to the entity and user to which they belong, along with their corresponding labels as illustrated in Figure 1. The generated samples are feature vectors containing a set of indicators that represent the behaviour of an entity associated with a specific user, identified by $user_id$, within a specified time range, defined by a timestamp called $time_{ini}$, with a fixed sample length of 1 hour as stated in [3]. Technically speaking, the RBD24 dataset encompasses 586,510 unique samples, from 1,772 unique users, with a global inventory of almost 1,386 workstations and 1,024 portable entities. Typically, each user has at least one workstation device and one portable device such a smartphone or a tablet device. In detail, the smartphones and tablet devices include both iOS and Android systems, whereas workstations span Linux, Windows, and macOS computers. Table I provides a summary of the risk activities included in the RBD24 dataset.

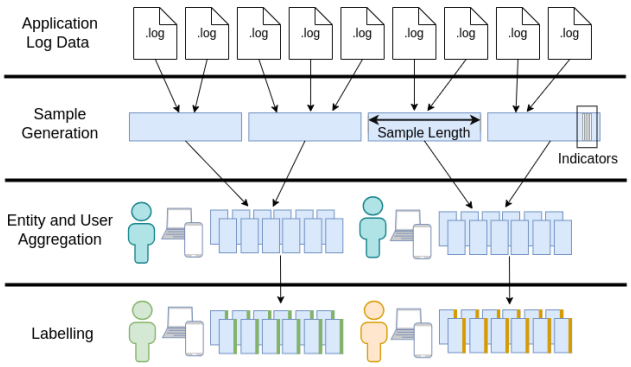


Fig. 1. Process of transforming application logs to labelled samples

III. Discussion & Conclusions

The RBD24 dataset represents a significant contribution to the cybersecurity research community by providing an open source dataset focused on risk activities detected through multi-modal application logs. This dataset facilitates the advancement of UEBA and machine learning-based risk assessment methodologies.

One of the key advantages of RBD24 is its emphasis on application logs, which capture high-level user activities beyond traditional network traces. This shift is crucial in modern cybersecurity, where attackers increasingly leverage application-layer vulnerabilities and sophisticated evasion techniques. By incorporating multiple log sources, RBD24 enables researchers to develop more robust detection strategies that reflect real-world attack behaviours. Furthermore, the structure and organisation of the dataset allow for scalable analysis, supporting investigations into the aforementioned cybersecurity threats. The ground truth labelling provided through intrusion detection systems and controlled experiments further enhances its reliability for supervised learning tasks.

In the future, we envision multiple research opportunities that stem from RBD24. The dataset can serve as a reference for evaluating anomaly detection models, as well as for developing early warning systems that integrate application log analytics. Moreover, its multi-modal nature paves the way for hybrid approaches that combine log data with endpoint telemetry and behavioural profiling.

RBD24 fills a critical gap in publicly available cybersecurity datasets by providing a labelled, real-world collection of risk activities extracted from application logs. Its release aims to drive innovation in cyber risk assessment and UEBA, fostering the development of more effective security solutions to combat emerging threats.

TABLE I
Summary of the RBD24 dataset, detailing risk activities collected from Desktop (DE) and Smartphone (SM) devices.

DatasetId	Risk Activity	Device	Benign (0) [#User/#Samples]	Risk (1) [#User/#Samples]
Crypto	Miner Checking	DE	738/161202	11/1343
		SM	613/180021	4/956
OutFlash	Outdated software	DE	738/161202	56/10820
		SM	613/180021	22/6639
OutTLS	Outdated TLS	DE	738/161202	18/2458
		SM	613/180021	11/2930
P2P	P2P Activity	DE	738/161202	177/35892
		SM	613/180021	94/21688
NonEnc	Non-encrypted	DE	738/161202	291/59943
		SM	613/180021	167/41434
Phishing	Phishing email	DE	98/13864	19/3072
		SM	117/34006	26/8968

Ethical Compliance

The data used in this study were collected in accordance with established ethical guidelines and were approved by an institutional ethics committee. All data are pseudo-anonymized at the time of collection, ensuring that no personally identifiable information is accessible.

Acknowledgement

This work received funding from CDTI CERVERA and was supported under Grant Agreement no. 101093164 ExtremeXP. The content of this article does not reflect the official opinion of the European Union, CDTI, or any other institution.

References

[1] A. Calvo Ibañez, S. Escuder, N. Ortiz, J. Escrig, and M. Compastié, "Rbd24: A Labelled Dataset with Risk Activities Using Log Application Data," Rochester, NY, Sep. 2024. [Online]. Available: https://papers.ssrn.com/abstract=4963547

[2] S. Garcia, A. Parmisano, and M. J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic," Jan. 2020. [Online]. Available: https://zenodo.org/records/4743746

[3] A. Calvo, S. Escuder, J. Escrig, M. Arias, N. Ortiz, and J. Guijarro, "A Data-driven Approach for Risk Exposure Analysis in Enterprise Security," in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. Thessaloniki, Greece: IEEE, Oct. 2023, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10302480

Sesión Pósters III: Aplicaciones de IA en ciberseguridad

A review of comparative analysis of natural language processing models for spam email identification

Francisco Jáñez-Martino*† francisco.janez@unileon.es
Eduardo Fidalgo*† eduardo.fidalgo@unileon.es
Rocío Alaiz-Rodríguez*† rocio.alaiz@unileon.es
Alicia Martínez-Mendoza*† alicia.martinez@unileon.es
Encina Díez-del-Río* mdiezr@unileon.es
*Universidad de León, Campus Vegazana, 24007 León, Spain
†Researcher at INCIBE

***Abstract*—Spam emails can be a major vector for cyberattacks, often containing malware embedded in links or attachments. Early and accurate detection of such emails is critical for cybersecurity, as it helps prevent system compromises and financial losses. This study evaluates state-of-the-art NLP models for binary classification of spam emails as malicious or non-malicious, comparing fastText, BERT, RoBERTa, DistilBERT, XLM-RoBERTa, and Large Language Models (LLaMA, Mistral) on the SEMD-600 dataset. The results indicate that DistilBERT achieves the best trade-off between performance and efficiency, with a recall of 0.792 and a runtime of 1.612 ms per email. These findings underscore the effectiveness of transformer-based models for malware spam detection, providing valuable insights for cybersecurity organizations seeking to enhance email filtering systems and strengthen early threat detection mechanisms.**

***Index Terms*—text classification, spam email detection, malware detection, cybersecurity**

Type of contribution: *Already published research [1]*

I. Introduction

Spam has long posed a significant cybersecurity challenge, leading to productivity losses, financial costs, and increased security risks. Beyond being a nuisance, it serves as a primary attack vector for scams and malware, including ransomware, which can compromise entire systems. Organizations face severe repercussions, such as data breaches, unauthorized access, and system takeovers, while individuals are vulnerable to confidential data theft and identity fraud.

Despite deploying anti-spam filters, many fail to accurately identify high-risk emails that embed malware within links or attachments. Spammers continuously adapt, employing evasive techniques such as word obfuscation, URL cloaking, and polymorphic attacks, making detection increasingly challenging. This adversarial evolution degrades datasets and machine learning models over time, requiring continuous adaptation and refinement of spam filtering strategies [2].

We evaluated eight Natural Language Processing (NLP) models, from traditional methods to Transformer-based models and large language models (LLMs), to detect spam emails with malware based on textual content, following Redondo-Gutierrez et al. [3]. Using the SEMD-600 dataset due to the lack of public alternatives, we compared model performance and provided recommendations for anti-spam filter implementation.

Malware detection has been explored using NLP techniques, with models such as Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) applied for classification [4]. Alam [5] used NLP for semantic similarities in malicious code, although applying these techniques to spam emails may increase analysis time due to URLs and attachments. While some studies focus on malware in files, this work targets spam emails containing malware. Abu Qbeitah and Aldwairi [6] studied dynamic anomaly detection in phishing emails, Cohen et al. [7] analyzed malware propagation in webmail attachments, and Arivudainambi et al. [8] used Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) for malware surveillance. Building on previous research [3], this study advances spam email detection through rapid textual analysis, comparing traditional methods such as (Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BoW) with Logistic Regression (LR), and Random Forest classifiers) with advanced models like Transformers and LLMs for improved performance.

II. Methodology

We followed a text classification approach to identify whether spam emails contain malware or not based only on textual content. We evaluated our proposal on the Spam Email Malware Detection 600 (SEMD-600) dataset [1], created by Redondo-Gutierrez et al. [3]. It consists of 300 spam emails with malware and 300 without malware, all written in English. The authors obtained the highest performance for spam email classification based on traditional techniques using the TF-IDF vectorizer and LR. For evaluation, we used their best result as the baseline. We compared it with other state-of-the-art models, including FastText word embedding, four Transformer-based models (BERT, RoBERTa, DistilBERT, and XLM-RoBERTa), and two LLMs (LLaMA and Mistral). Thereby, we provided new baseline results for the task of spam malware detection using only the text of the email.

III. Experimentation settings and results

We conducted the experiments on a computer with 128 GB of RAM, two Intel Xeon E5-2630v3 processors of 2.4 GHz, and two Nvidia Titan Xp. We used Python packages to prepare the implementation, training, and evaluation of the models (simpletransformers, transformers, and fastText).

For BERT, RoBERTa, DistilBERT, XLM-RoBERTa, and LLMs, a maximum of 512 tokens was configured, along with a learning rate of 0.00001 and 10 training epochs using 8 as the training batch. FastText was trained from scratch with 200 epochs.

[1]https://gvis.unileon.es/datasets-semd-600/ retrieved March 2025

The models were evaluated in terms of precision, recall, F1-score, accuracy, and execution time per email, and we used 5-fold cross-validation due to the small dataset size.

Our main goal is to maximize recall, ensuring the detection of as many malware-infected spam emails as possible. The results presented in Table I indicate that DistilBERT and RoBERTa achieved the highest recall values. However, DistilBERT demonstrated superior precision and F1-score, making it the most suitable model for this task.

The traditional TF-IDF with Logistic Regression (LR) model exhibited higher recall than BERT and XLM-RoBERTa, suggesting that simpler models may be more effective in specific spam detection scenarios. Conversely, FastText was the worst-performing model, likely due to its reliance on word embeddings, which may fail to capture the intricate patterns and obfuscation techniques used in spam emails.

Regarding Large Language Models (LLMs), both LLaMA and Mistral yielded lower overall performance, with higher precision than recall. This suggests that these models may excel at identifying emails with strong malware indicators but struggle with borderline cases, misclassifying some malware-containing emails as harmless.

TABLE I
EVALUATION OF BASELINE RESULTS (TF-IDF-LR) FOR SPAM MALWARE DETECTION AGAINST THE ONE-WORD EMBEDDING MODEL (FASTTEXT), FOUR ATTENTION MODELS, AND TWO LLMS IN TERMS OF PRECISION (P), RECALL (R), F1-SCORE (F1), AND ACCURACY (ACC).

Model	P	R	F1	Acc (%)
TF-IDF-LR	0.768	0.763	0.763	76.4
FastText	0.730	0.643	0.681	68.7
BERT	0.733	0.734	0.730	71.9
RoBERTa	0.743	**0.792**	0.766	74.8
DistilBERT	0.774	**0.792**	0.781	77.0
XLM-RoBERTa	0.718	0.780	0.746	72.4
LLaMA	0.620	0.594	0.606	59.8
Mistral	0.653	0.593	0.621	62.4

The execution time was also analyzed, as spam email detection represents a big data challenge, where detection speed is a critical factor. Among the models evaluated, FastText and the traditional TF-IDF-LR model exhibited the fastest processing times, with execution speeds of 0.164 ms and 0.278 ms per email, respectively.

Among attention-based models, DistilBERT demonstrated the best trade-off between efficiency and accuracy, achieving an execution time of 1.612 ms per email while maintaining high recall and F1-score. This makes DistilBERT the most suitable transformer-based model for malware detection in spam emails, as it offers a balanced combination of detection performance and computational efficiency.

The LLMs had significantly longer processing times (200.12 ms for LLaMA and 88.28 ms for Mistral), making them less suitable for real-time detection tasks.

IV. CONCLUSIONS

This study evaluated various NLP models for detecting malware-infected spam emails, formulating the task as a binary classification problem (malware vs. non-malware). The models, including both traditional and transformer-based approaches, were trained on the SEMD-600 dataset. The results indicate that DistilBERT achieved the best balance between recall and runtime efficiency. However, the overall recall remains below 80%, highlighting opportunities for further improvement. Additionally, the findings suggest that traditional models, such as TF-IDF with Logistic Regression, remain viable alternatives due to their ease of implementation and faster inference times. Regarding Large Language Models (LLMs), both LLaMA and Mistral did not perform well in this task.

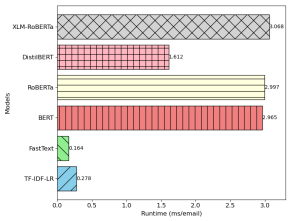


Fig. 1. Evaluation of the models in terms of execution times. The results are in milliseconds (ms) per email.

The relevance of cybersecurity in this study is underscored by its focus on the early detection of malicious emails, a primary attack vector for cybercriminals. By comparing different models, this research provides valuable insights into optimizing spam filtering techniques, ensuring effective threat detection without compromising processing speed.

ACKNOWLEDGEMENTS

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation) thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

REFERENCES

[1] F. Jáñez Martino, E. Fidalgo, R. Alaiz-Rodríguez, A. Carofilis, and A. Martínez-Mendoza, "Comparative analysis of natural language processing models for malware spam email identification," in *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2024)*, Lancaster, United Kingdom, 2024. [Online]. Available: https://nlpaics-2024.org/programme-2/

[2] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: analysis of spammer strategies and the dataset shift problem," *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1145–1173, 2023.

[3] L. A. Redondo-Gutierrez, F. Jáñez Martino, E. Fidalgo, E. Alegre, V. González-Castro, and R. Alaiz-Rodríguez, "Detecting malware using text documents extracted from spam email through machine learning," in *Proceedings of the 22nd ACM Symposium on Document Engineering*, ser. DocEng '22, 2022.

[4] R. Mehta, O. Jurečková, and M. Stamp, "A natural language processing approach to malware classification," *Journal of Computer Virology and Hacking Techniques*, vol. 20, no. 1, pp. 173–184, 2024.

[5] S. Alam, "Applying natural language processing for detecting malicious patterns in android applications," *Forensic Science International: Digital Investigation*, vol. 39, p. 301270, 2021.

[6] M. A. Qbeitah and M. Aldwairi, "Dynamic malware analysis of phishing emails," in *2018 9th International Conference on Information and Communication Systems (ICICS)*, 2018, pp. 18–24.

[7] Y. Cohen, D. Hendler, and A. Rubin, "Detection of malicious webmail attachments based on propagation patterns," *Knowledge-Based Systems*, vol. 141, pp. 67–79, 2018.

[8] D. Arivudainambi, K. A. Varun, S. Chakkaravarthy, and P. Visu, "Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance," *Computer Communications*, vol. 147, 08 2019.

A review of Enhancing CSIRT effectiveness with transformer-based cyber incident classification

Alicia Martínez-Mendoza* alicia.martinez@unileon.es
Eduardo Fidalgo* eduardo.fidalgo@unileon.es
Enrique Alegre* enrique.alegre@unileon.es
Waqar Tanveer* wtan@unileon.es
Encina Díez-del-Río* mdiezr@unileon.es
*Universidad de León, Campus Vegazana, 24007 León, Spain

***Abstract*—The increasing volume and complexity of cyber incidents necessitate efficient classification methods to enhance response times in Computer Security Incident Response Teams. This paper introduces a transformer-based classifier that categorizes cyber incident reports according to the taxonomy of INCIBE. A key contribution is the creation of the CECILIA dataset, comprising 923 manually labeled cyber incident reports sourced from six authoritative web platforms. To mitigate class imbalance, a reduced version, CECILIA-6C-900, consolidates low-frequency categories. Comparative evaluations of traditional classifiers and six transformer-based models reveal that XLNet achieves the highest performance, with an accuracy of 83.85% and an MCC of 0.7668. The paper highlights the potential of transformers architecture in cyber incident classification, enabling automated triage and faster response times for CSIRTs. The proposed dataset and baseline results establish a foundation for future research in AI-driven cybersecurity automation.**

***Index Terms*—transformer-based models, cyber incident classification, cybersecurity**

Type of contribution: *Already published research [1]*

I. Introduction

The rise in cyberattacks requires Computer Security Incident Response Teams (CSIRTs) to improve incident management, particularly in initial classification, as accurate categorization ensures incidents are routed to the right experts, enhancing response times [1]. The traditional approach to automatic classification relies on the standardization of incident reports. However, harmonizing cyber incident reporting is challenging due to the diversity of formats and standards. To address this issue, we propose leveraging Natural Language Processing (NLP)-based classification techniques[2], [3], [4], specifically transformer models, as no prior studies have applied these models to classify cyber incidents according to the INCIBE taxonomy. Previous research includes AI-assisted classification using OSINT and regression algorithms for risk-based categorization[5], NLP algorithms for predicting cyber threats in online discussions[6], and text classification studies where BERT and XLNet outperformed other models[7].

The paper introduces a novel dataset family, CECILIA (**C**yb**E**rin**CI**dents c**L**assified **I**ncibe t**A**xonomy), containing reports collected from six selected web sources and manually labeled according to the National Cybersecurity Institute (INCIBE) taxonomy[1]. We introduce two versions of the dataset: CECILIA-10C-900, which categorizes cyber incidents into ten distinct classes, and CECILIA-6C-900, a reduced version that consolidates underrepresented categories into a single class to address class imbalance. This study aims to systematically evaluate the effectiveness of different classification approaches in cyber incident categorization. To this end, we compare the performance of two traditional machine learning models and six transformer-based architectures, assessing their capability to enhance automated cyber incident classification. This comparative analysis provides insights into the most suitable methodologies for improving cyber defense mechanisms and optimizing incident response workflows within CSIRTs.

[1]https://www.incibe.es/incibe-cert/incidentes/taxonomia

II. CECILIA dataset

The CECILIA-10C-900 dataset consists of 923 cyber incident reports from six sources manually labeled by a cybersecurity expert and a labelling assistant into 10 classes and 38 subclasses according to the INCIBE taxonomy. The 10 classes are: abusive content (AC), malicious code (MC), information gathering (IG), intrusion attempts (IA), intrusions (I), availability (A), information content security (ICS), fraud (F), vulnerable (V), and others (O). This dataset is imbalanced, with the ICS category representing nearly 50% of the cases. An alternative dataset, CECILIA-6C-900, was created. This new version combines the four least representative categories (V, AC, IG and IA) into a single category labeled “others” (O), thereby mitigating the imbalance.

III. Evaluation of Transformers architecture

A. Experimental settings

The Simpletransformers library was used to compare transformer models under uniform conditions. Six transformer-based models were evaluated: DistilBERT, RoBERTa, XLNet, ELECTRA, Longformer and MPNet [7]. Additionally, baseline results were calculated using two traditional machine learning classifiers that had previously performed well in cybersecurity text classification problems: Logistic Regression with TF-IDF and K-Nearest Neighbor (KNN) with Bag of Words (BoW)[8].

Given the limited size of the dataset (923 samples), stratified cross-validation with $k = 5$ was used. The evaluation metrics included Accuracy, Variance, Precision, Recall, F1-score (the last three both weighted (w) and macro (m)), and lastly Matthews Correlation Coefficient (MCC).

B. Results and Discussion

The transformer-based models consistently outperformed traditional classifiers across all evaluation metrics. In the

TABLE I
INCIDENT CLASSIFICATION RESULTS OVER CECILIA-6C-900 DATASET, AFTER MERGING THE FOUR REPRESENTATIVE CATEGORIES, WITH TWO TRADITIONAL CLASSIFIERS (LR+TF-IDF AND KNN+BOW) AND SIX TRANSFORMER-BASED MODELS. THE BEST RESULTS ARE IN BOLD.

Model	Accuracy	Variance	Prec-w	Prec-m	Recall-w	Recall-m	F1-score-w	F1-score-m	MCC
LR TF-IDF	0.7292	0.0000	0.7293	0.6665	0.7292	0.6024	0.7268	0.6252	0.6048
KNN BoW	0.5848	0.0000	0.5267	0.3881	0.5848	0.3372	0.5290	0.3255	0.3558
DiltilBERT	0.8093	0.0007	0.7994	0.7305	0.8093	0.7093	0.8013	0.7130	0.7205
RoBERTa	0.8169	0.0003	0.8186	0.7384	0.8169	0.7384	0.8151	0.7338	0.7350
XLNet	0.8266	0.0006	0.8300	0.7492	0.8231	0.7272	0.8265	0.7476	0.7490
ELECTRA	0.7865	0.0002	0.7699	0.6436	0.7865	0.6000	0.7667	0.5919	0.6887
Longformer	0.8201	0.0003	0.8166	0.7453	0.8201	0.7538	0.8167	0.7465	0.7399
MPNet	**0.8352**	0.0007	**0.8314**	**0.7741**	**0.8352**	**0.7489**	**0.8273**	**0.7494**	**0.7608**

CECILIA-10C-900 dataset, XLNet achieved the highest performance, with an accuracy of 0.8385 and an MCC of 0.7668, followed by RoBERTa, which obtained 0.8245 accuracy and 0.7463 MCC. Although ELECTRA demonstrated the lowest performance among the transformer-based models (0.7984 accuracy and 0.7059 MCC), it still surpassed traditional classifiers, reinforcing the advantages of deep learning approaches in this domain. The superior results of XLNet and RoBERTa highlight the importance of contextual understanding in cyber incident classification, as these models effectively capture linguistic patterns and relationships essential for accurately categorizing incidents. MPNet did not achieve a performance comparable to BERT and XLNet. This deviation may be attributed to the limited dataset size and the relatively short length of incident reports, which could hinder the model's ability to leverage its full potential. Among the models optimized for computational efficiency, DistilBERT demonstrated the best trade-off between speed and accuracy, achieving the lowest inference time of 0.79 seconds per sample, making it a suitable candidate for real-time cyber incident classification in resource-constrained environments.

C. Dataset Balancing

To address the imbalance in the CECILIA-10C-900 dataset, the four least represented categories (AC, V, IG, IA) were grouped into a new category named "others" (O), resulting in the CECILIA-6C-900 dataset. The evaluation of the classification models on this dataset is presented in Table I. The best-performing model is MPNet, with an accuracy of 0.8352 and 0.7608 MCC. Furthermore, the F1-score values for each cross-validation split for the less frequent categories ranged from 0.27 to 0.45, while for the CECILIA-10C-900 dataset, some classes were never correctly classified. Therefore, balancing the dataset enhances the model's overall performance.

IV. CONCLUSIONS

This study evaluated the performance of two traditional classifiers and six transformer-based models using the proposed CECILIA datasets. The results demonstrated that transformer-based architectures consistently outperformed traditional machine-learning approaches across all evaluated metrics. XLNet and RoBERTa exhibited superior performance in cyber incident classification, achieving accuracy values of 0.8385 and 0.8245, respectively. However, these models encountered challenges in accurately classifying underrepresented categories, primarily due to class imbalance within the dataset. The CECILIA-6C-900 dataset was introduced to mitigate this issue, consolidating the four least frequent categories into a single class. MPNet yielded the highest performance on this balanced dataset, attaining an accuracy of 0.8353. These findings highlight two key factors for improving cyber incident classification: (i) the effectiveness of transformer models in capturing contextual dependencies and (ii) the importance of addressing class imbalance in cybersecurity datasets. These insights may have significant implications for CSIRTs to enhance automated cyber incident processing and develope more efficient and responsive cybersecurity defense strategies.

ACKNOWLEDGMENTS

This work was supported by the Strategic Project LUCIA granted to the University of León by the Spanish National Cybersecurity Institute (INCIBE) and funded by the Next Generation EU funds.

REFERENCES

[1] J. J. Delgado, E. Fidalgo, E. Alegre, A. Carofilis, and A. Martínez-Mendoza, "CECILIA: Enhancing CSIRT Effectiveness with Transformer-Based Cyber Incident Classification," in *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2024)*, Lancaster, United Kingdom, 2024. [Online]. Available: https://nlpaics-2024.org/programme-2/

[2] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and R. Aláiz-Rodríguez, "File name classification approach to identify child sexual abuse," in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, vol. 1, no. 978-989-758-397-1, 2020.

[3] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and R. Alaiz-Rodriguez, "Short text classification approach to identify child sexual exploitation material," *Scientific reports*, vol. 13, no. 1, p. 16108, 2023.

[4] L. Á. Redondo-Gutierrez, F. Jáñez-Martino, E. Fidalgo, E. Alegre, V. González-Castro, and R. Alaiz-Rodríguez, "Detecting malware using text documents extracted from spam email through machine learning," in *Proceedings of the 22nd ACM Symposium on Document Engineering*, 2022, pp. 1–4.

[5] A. Tundis, S. Ruppert, and M. Mühlhäuser, "A Feature-driven Method for Automating the Assessment of OSINT Cyber Threat Sources," *Computers & Security*, vol. 113, p. 102576, feb 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167404821004004

[6] A. Sapienza, S. K. Ernala, A. Bessi, K. Lerman, and E. Ferrara, "Discover: Mining online chatter for emerging cyber threats," in *Companion Proceedings of the The Web Conference 2018*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 983–990. [Online]. Available: https://doi.org/10.1145/3184558.3191528

[7] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A Survey on Text Classification Algorithms: From Text to Predictions," *Information 2022, Vol. 13, Page 83*, vol. 13, no. 2, p. 83, feb 2022. [Online]. Available: https://www.mdpi.com/2078-2489/13/2/83/htmhttps://www.mdpi.com/2078-2489/13/2/83

[8] L. Á. Redondo-Gutierrez, F. Jáñez-Martino, E. Fidalgo, E. Alegre, V. González-Castro, and R. Alaiz-Rodríguez, "Detecting malware using text documents extracted from spam email through machine learning," *DocEng 2022 - Proceedings of the 2022 ACM Symposium on Document Engineering*, sep 2022.

A review of Towards Multi-Class Smishing Detection: A Novel Feature Vector Approach and the Smishing-4C Dataset

Alicia Martínez-Mendoza*† alicia.martinez@unileon.es
Francisco Jáñez-Martino*† francisco.janez@unileon.es
Laura Fernández-Robles*‡ l.fernandez@unileon.es
Enrique Alegre*† enrique.alegre@unileon.es
Eduardo Fidalgo*† eduardo.fidalgo@unileon.es
Irene Aguado-Caballero*† iaguc@unileon.es

*Department of Electrical, Systems and Automation Engineering, Universidad de León, Campus Vegazana, 24007 León, Spain
‡Department of Mechanical, Informatics and Aerospace Engineering, Universidad de León, Campus Vegazana, 24007 León, Spain
†Researcher at INCIBE (Spanish National Cybersecurity Institute)

***Abstract*—Smishing uses phishing techniques to target Short Message Service users. The attacker disseminates fraudulent messages and impersonates a legitimate entity to steal personal information and credentials. In our paper , we introduced Smishing-4C, a dataset containing the four most common types of smishing: Bank/Finance, Rewards, Dating, and SMS. In addition, we created a six-feature vector that was applied to the smishing classification task, leading to remarkable results. Bag of Words and the proposed six-feature vector outperformed transformer-based models on Smishing-4C, obtaining an F1 score of 0.788. The multi-class classification provides information on how attackers reach people, which can be useful for ensuring cybersecurity measures in SMS services.**

***Index Terms*—cybersecurity, text classification, smishing classification, multiclass classification, Smishing-4C dataset**

Type of contribution: *Already published research [1]*

I. Introduction

Smishing is a phishing technique in which an attacker uses the Short Message Service (SMS) to deliver a phishing attack. Most authors focused on binary classification rather than distinguishing between different types of scams [1]. A multiclass perspective allows the classification of social engineering techniques more easily since more information is analyzed [2]. This would allow us to identify target users and risks more efficiently. Thus, it will help to leap forward in protecting against cybersecurity fraud. Therefore, we proposed the following contributions:

- Smishing-4C, a text-based smishing dataset. It is labeled in four categories of smishing used for the first time in multi-class smishing classification.
- A novel feature vector with six features based on text, which had not been used before in smishing detection or classification.
- Evaluation of our dataset and feature vector with four machine learning and deep learning models.

II. Related work

Previous research on smishing detection reveals that authors mainly have studied classification into legitimate and fraudulent messages. Some authors have adopted additional stages in the detection process. Jain et al. [3] and Alegre et al. [4] adopted phishing URL detection and classification, and Jáñez-Martino et al. [5] natural language processing (NLP) malware identification. Others like Timko and Rahman [6] or Sonowal [7] have developed more advanced methods such as vectors with smishing features like URLs or phone numbers.

Timko and Rahman [6] have conducted a more detailed analysis of smishing, thereby identifying different categories, although they have not used them for classification.

Regarding short-text classification, Al-Nabki et al. [8] ilustrated that traditional models can outperform transformer-based models due to the lack of context in unstructured texts. Thus, we included a comparison between machine and deep learning models.

III. Methodology

A. Smishing-4C

We created Smishing 4-C, the first dataset in the literature that introduced a multiclass approach to identify the most common types of fraud employed in English smishing attacks. Smishing-4C combines samples from two publicly available datasets: Kaggle Smishing [9] and Mendeley Smishing [10].

Smishing-4C contains 120 smishing texts in English, manually labeled for smishing classification. The dataset is balanced, and the definitions of the classes are given below:
Bank/Finance: messages from a bank or financial services entity.
SMS Service: notifications regarding unread messages, voicemails, subscriptions to a service, or customer service announcements.
Dating: messages related to dating, friendship, personal relationships, secret admirer messages, or sexual content.
Rewards: messages indicating the user has received an award, a free product, or a prize.

B. Smishing features

To build the proposed six-feature vector, we selected the following features from previous works: Length value, Number of writing errors, Phone, and URL [6], [7]. Additionally, we included two other important features for smishing detection: Company names and Slang. Each feature

is described below:
Length value: number of characters in the SMS.
Number of writing errors: number of writing errors found in the SMS.
Phone: indicates if the message contains a phone number.
URL: indicates whether the message contains a URL, making a distinction between long and shortened URL.
Company name: contains the string corresponding to a company name present on the SMS.
Slang: indicates whether the SMS contains slang.

C. Evaluation method

To validate the performance of the proposed six-feature vector, we selected four machine learning models: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). We also utilized Bag of Words (BOW) and N-grams as vectorizers of the models to incorporate methods that can better understand the context of the messages. We compared their performance with different types of neural networks: a basic Multilayer Perceptron (MLP), a Long Short-Term Memory (LSTM), and two transformer models, Bidirectional Encoder Representations from Transformers (BERT), and Enhanced Representation through Knowledge Integration (ERNIE).

IV. Results

The transformer model that obtained the highest F1-score of 0.701 on the Smishing-4C dataset was ERNIE. As far as the machine learning models are concerned, the best results were obtained with LR. BoW yielded an F1-score of 0.691, whereas the combination of BoW with the six-feature vector reached an F1-score of 0.764, demonstrating that the smishing features enhance the differentiation of smishing classes. Furthermore, an enhancement in the performance of the other classifiers (SVM, RF and DT) is observed when the feature vector is incorporated. Results of our approach are shown in Table I.

We observed that SMS service is the category with the lowest F1-score in both ERNIE and LR models, while Bank/Finance and Dating obtained a higher performance. In particular, we observed that the most favorable results were obtained for the Bank/Finance class, yielding a 0.652 F1-score using only the feature vector. The other classes presented a more significant challenge, and the optimal performance was achieved by integrating the text-based information (BoW) with the six-feature vector.

TABLE I
PERFORMANCE RESULTS FOR THE MACHINE AND DEEP LEARNING MODELS ON THE SMISHING-4C DATASET. THE LAST ROWS (BOW + SIX FEATURES) SHOW HOW OUR APPROACH IMPROVES THE PERFORMANCE OF THE MODELS.

Feature Extraction	Classifiers	Precision	Recall	F1-Score
BERT	BERT	0.731	0.692	0.700
ERNIE	ERNIE	0.706	0.700	0.701
LSTM	LSTM	0.682	0.667	0.670
BoW	SVM	0.700	0.650	0.648
	LR	0.724	0.696	0.691
	RF	0.736	0.681	0.679
	DT	0.715	0.654	0.653
BoW + six features	SVM	**0.773**	0.758	0.758
	LR	0.772	**0.767**	**0.764**
	RF	0.767	0.747	0.746
	DT	0.718	0.697	0.692

V. Conclusions

In this paper, we have presented Smishing-4C, a dataset labeled for four classes of smishing: Bank/Finance, Rewards, Dating, and SMS Service. We have evaluated four traditional and four deep learning models on Smishing-4C. To increase the performance of the traditional models, we proposed a six-feature vector, and we evaluated its combination with BoW and (1,4)-gram representation. Results showed that the six-feature vector and BoW combination achieved an F1-score of 0.764, outperforming transformers models. This could be very beneficial, since transformers consume more resources than traditional methods, and the classification could be done faster and more efficiently.

The automatic classification of smishing types can enhance the security of the SMS environment. It would be a useful tool for Computer Emergency Response Teams to automatically process smishing reports. It would also enable the distribution of alerts to the type of entities identified within the smishing categories, such as banking or financial institutions. These alerts could be customized to target specific users within these entities, contributing to reduce the number of victims.

Acknowledgments

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation), thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

References

[1] Martínez-Mendoza, A., Jáñez-Martino, F., Carofilis, A., Fernández-Robles, L., Alegre, E., & Fidalgo, E. (2024). "Towards Multi-Class Smishing Detection: A Novel Feature Vector Approach and the Smishing-4C Dataset", XL Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), 3846, 58-68.

[2] Jáñez-Martino, F. (2024). "Analysis and classification of spam email using Artificial Intelligence to identify cyberthreats". In Procesamiento del Lenguaje Natural, 72, 155–158.

[3] Jain, A. K., Gupta, B. B., Kaur, K., Bhutani, P., Alhalabi, W., & Almomani, A. (2022). "A content and URL analysis-based efficient approach to detect smishing SMS in intelligent systems". International Journal of Intelligent Systems, 37(12), 11117-11141.

[4] Alegre, E., Fidalgo, E., Fernandez-Robles, L., Martínez-Mendoza, A., & Biswas, R. (2023). "Smishing Detection: Recognition of URLs from SMS Screenshots". In 5th International Conference on Applications of Intelligent Systems (Las Palmas de Gran Canaria, España).

[5] Jáñez-Martino, F., Fidalgo, E., Alaiz-Rodríguez, R., Carofilis-Vasco, A., & Martínez-Mendoza, A. (2024). "Comparative Analysis of Natural Language Processing Models for Malware Spam Email Identification". In International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS 2024).

[6] Timko, D., & Rahman, M. L. (2023). "Commercial anti-smishing tools and their comparative effectiveness against modern threats". In Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks (pp. 1-12).

[7] Sonowal, G. (2020). "Detecting phishing SMS based on multiple correlation algorithms". SN computer science, 1(6), 361.

[8] Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Alaiz-Rodríguez, R. (2023). "Short text classification approach to identify child sexual exploitation material". Scientific reports, 13(1), 16108.

[9] Kaggle. (2022). SMS Smishing collection dataset. Retrieved from https://www.kaggle.com/datasets/galactus007/sms-smishing-collection-data-set

[10] Mishra, S., & Soni, D. (2022). "SMS Phishing Dataset for Machine Learning and Pattern Recognition". In International Conference on Soft Computing and Pattern Recognition (pp. 597-604).

Resumen de detección avanzada de smishing utilizando el dataset IRIS y visión artificial

Pablo Blanco-Medina*† pablo.blanco@unileon.es
Andrés Carofilis‡ andres.carofilis@idiap.ch
Eduardo Fidalgo*† eduardo.fidalgo@unileon.es
Enrique Alegre*† enrique.alegre@unileon.es
Irene Aguado-Caballero* iaguc@unileon.es

*Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León, Campus Vegazana, 24007 León, Spain
†Investigador colaborador con INCIBE (Spanish National Cybersecurity Institute)
‡Idiap Research Institute, Martigny, Switzerland

***Resumen*—Los Equipos de Respuesta ante Emergencias Informáticas (CERTs) reciben capturas de pantalla que podrían tener smishing, que analizan manualmente. Por este motivo, se propone: (1) un modelo para clasificar automáticamente imágenes que podrían contener smishing y un dataset (IRIS-244) para su ajuste, donde la arquitectura Xception obtiene los mejores resultados (78,36 % de precisión); (2) la extracción automática de URLs (Localizador de recursos uniforme) de mensajes de smishing usando Text Spotting, mejorando la precisión del reconocimiento del 3,05 % al 22,90 %; y (3) la inclusión de preprocesado y OCRs en dicha extracción, para llegar a una precisión final del 61,16 % en la recuperación de URLs. La automatización del análisis de mensajes de smishing mediante el uso de herramientas que realizan la extracción de información de forma automática es beneficiosa para los CERTs, facilitando la detección temprana de campañas maliciosas.**

***Index Terms*—Ciberseguridad, Seguridad, Aprendizaje Profundo, Apoyo a Operadores Humanos, Redes Sociales, Smishing.**

Tipo de contribución: *Investigación ya publicada [1], [2], [3]*

I. Introducción

La mensajería por SMS (Servicio de mensajería corta) es usada como vector de ataque en campañas de smishing, donde los ciberdelincuentes envían mensajes fraudulentos que suplantan la identidad de entidades legítimas para usurpar información mediante enlaces maliciosos [4]. Los CERTs reciben capturas de pantalla de estos ciberdelitos diariamente, por lo que requieren de herramientas automatizadas para procesar estas campañas con rapidez y eficiencia. En los trabajos desarrollados buscábamos procesar capturas de pantalla de posibles mensajes de smishing para clasificarlas según si el mensaje estaba dividido [1] y recuperar su URL automáticamente[2], [3]. Para ello, propusimos: (i) aplicar redes neuronales y Transformers para clasificar nuestro conjunto de capturas de imágenes Smishing IRIS-244 (sm**I**shing sc**R**eenshots w**I**th url**S**) [1], (ii) aplicar técnicas de preprocesado, operaciones morfológicas y algoritmos de reconocedores ópticos de caracteres (OCR) para recuperar URLs en SMS[2], y (iii) la detección de URLs mediante Text Spotting y una metodología para unir componentes de las URLs[3], [5].

II. Trabajos relacionados

Tradicionalmente en clasificación de imágenes, se usaban redes neuronales para clasificar imágenes etiquetadas manualmente, preentrenandolas en grandes conjuntos como ImageNet [6]. En nuestro caso, estudiamos la tarea de clasificación de imágenes con texto y URLs, para lo que algunos autores han propuesto el uso de OCRs [7]. Estos modelos son útiles en su aplicación a capturas de pantalla con smishing. Otras propuestas significativas para la resolución de estos problemas son el uso de Aprendizaje Profundo para la categorización de campañas de smishing [5] o la aportación de datasets actualizados [8].

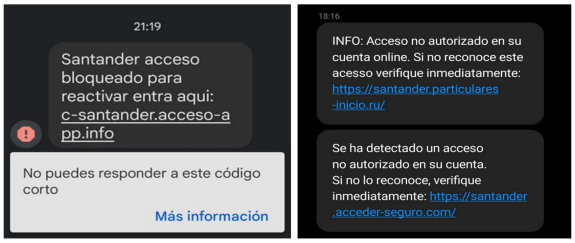


Figura 1. Imágenes de nuestro conjunto de datos IRIS-244, mostrando URLs divididas en múltiples líneas

III. Metodología

IRIS-244 (sm**I**shing sc**R**eenshots w**I**th url**S**) es un conjunto de datos de 244 imágenes de smishing, recopiladas con ayuda del INCIBE, que contienen 262 URLs repartidas en 117 imágenes con URLs divididas en varias líneas y 127 imágenes con URLs completas en una sola. La figura 1 presenta algunas de estas imágenes.

Para identificar la mejor arquitectura en el problema de detección de Smishing, seleccionamos (i) Xception [9] y ConvNext [10], (ii) un Vision Transformer [11] con capas Multi-Atención, (iii) una CNN con capas de reescalado y (iv) redes preentrenadas, comparando su rendimiento.

Para obtener las URLs de las imágenes, usamos el OCR Pytesseract [12]. Aplicamos diversos preprocesados a las imágenes, como la conversion de RGB a escala de grises y thresholding. Extraído el texto, comprobamos si la URL documentada se había recuperado correctamente, obteniendo la precisión final del método. Finalmente, evaluamos nuestra propuesta en 117 imágenes obtenidas de INCIBE, con mensajes que podrían contener enlaces maliciosos de Phishing.

Tabla I
PRECISIÓN DE LOS MODELOS EN CADA PARTICIÓN AL EVALUAR EN EL CONJUNTO IRIS-244

Fold	Transformer	Red Neuronal	Xception	ConvNextT
F1	53.06 %	44.90 %	71.43 %	46.94 %
F2	61.22 %	53.06 %	61.22 %	61.22 %
F3	61.22 %	55.10 %	75.51 %	73.47 %
F4	55.10 %	51.02 %	85.71 %	73.47 %
F5	58.33 %	62.50 %	**97.92 %**	91.67 %
Media	57.79 %	53.32 %	**78.36 %**	69.35 %

Tabla II
RENDIMIENTO EN PRECISIÓN DE LA RECUPERACIÓN DE URLS

Método	URL-C	URL-D	Total
PAN++	6,61 %	0,00 %	3,05 %
Mejoras Propuestas	25,62 %	20,57 %	22,90 %

En la segunda prueba para la detección y reconstrucción de URLs, usamos 244 imágenes de Smishing recopiladas por INCIBE. Para cada imagen, etiquetamos la transcripción del texto y la localización espacial de sus URLs. Aplicamos Text Spotting, comprobamos aquellas regiones con texto resaltado y las unimos en una sola, formando la URL final y comparándola con la documentada. Para recuperar el texto presente en estas imágenes, seleccionamos el método Pixel Aggregation Network (PAN++) [13], pre-entrenado en datasets de detección y reconocimiento de texto. Comprobamos si las regiones detectadas correspondían con las URLs etiquetadas y propusimos como candidatas las regiones subrayadas o con un color distinto al del texto. Usando umbralización y escala de grises, recuperamos las regiones detectadas por PAN++ y devolvimos las URLs del mensaje.

IV. EXPERIMENTACIÓN

Para evaluar los modelos de detección de Smishing utilizamos validación cruzada con 5 particiones, cuyos resultados pueden verse en la Tabla I. Estos resultados muestran que Xception fue la arquitectura con mejor rendimiento, con una precisión media de $78{,}36\,\%$, con su máximo rendimiento en el quinto fold ($97{,}92\,\%$), y seguida por ConvNextT, con $69{,}35\,\%$. Esto sugiere que Xception podría ser particularmente efectiva para la clasificación de capturas de pantalla de smishing.

En el experimento de detección de URLs, la configuración que mejores resultados obtuvo fue el preprocesado en escala de grises, con un $61,16\,\%$ de precisión. Tras realizar una revisión manual de las URLs no recuperadas, identificamos los dos errores más comunes como: (i) omisión de zonas de texto detectadas, y (ii) confusión entre caracteres similares.

En el tercer experimento, los resultados de la Tabla II muestran que PAN++ logró recuperar correctamente un $6{,}61\,\%$ de las URLs completas y no pudo recuperar ninguna URL separada ($0{,}00\,\%$), con un rendimiento total de $3{,}05\,\%$. Al incorporar preprocesamiento y lógica de unión de fragmentos de URLs, la precisión mejoró notoriamente en ambas categorías, aumentando del $6{,}61\,\%$ a $25{,}62\,\%$ y del $0{,}00\,\%$ a $20{,}57\,\%$ en URLs completas y divididas, respectivamente.

V. CONCLUSIONES

En este trabajo se ha abordado el problema de la recuperación de texto y URLs en el contexto de campañas de Smishing, por su importancia en la ciberseguridad y asistencia a equipos CERTs. Para ello, hemos desarrollado un clasificador de imágenes de Smishing para identificar la distribución de texto y hemos generado el conjunto de datos IRIS-244, con el que hemos medido el rendimiento de diversas arquitecturas concluyendo que Xception, con una precisión media de $78,36\,\%$, es la más adecuada a este problema.

Para recuperar e identificar correctamente el texto de las URLs, hemos combinado la aplicación del OCR Pytesseract con distintos tipos de preprocesado, mejorando la calidad del texto y evaluando las propuestas. Obtuvimos una mejor puntuación de $61,16\,\%$ de precisión mediante el preprocesado en escala de grises sobre el resto de operaciones de mejora.

Por último, buscamos extraer la URL contenida en el mensaje. Usando un método de Text Spotting, PAN++, reconstruimos la URL original y evaluamos nuestra propuesta en el conjunto IRIS-244. Obtuvimos una precisión de $22,90\,\%$ en el reconocimiento de URLs, mejorando la precisión del método original que era de un $3,05\,\%$.

AGRADECIMIENTOS

Este trabajo ha sido financiado gracias al Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation) gracias al Proyecto LUCIA (Lucha contra el Cibercrimen utilizando Inteligencia Artificial) concedido por INCIBE a la Universidad de León.

REFERENCIAS

[1] Blanco-Medina, P., Carofilis, A., Fidalgo, E. & Alegre, E. 2024. "Smishing Screenshots Classification using Deep Learning and IRIS Dataset" Jornadas de Automática, 45.

[2] Blanco-Medina, P., Carofilis, A., Fidalgo, E. & Alegre, E. 2024. "Enhancing CERT Smishing Response through Automatic SMS Screenshot URL Extraction." Jornadas de Automática, 45.

[3] Blanco-Medina, P., Biswas, R., Gonzalez-Castro, V., Alaiz Rodríguez, R., Fidalgo, E. & Alegre, E. 2024. "Enhancing Smishing URL Extraction with Text Spotting." Jornadas de Automática, 45.

[4] Jañez-Martino, F., Alaiz-Rodríguez, R., Gonzalez-Castro, V., Fidalgo, E. & Alegre, E., 2023. "A review of spam email detection: analysis of spammer strategies and the dataset shift problem." Artificial Intelligence Review 56 (2), 1145–1173.

[5] Blanco-Medina, P., Fidalgo, E., Alegre, E. & Gonzalez-Castro, V., 2022. "A survey on methods, datasets and implementations for scene text spotting." IET Image Processing 16 (13), 3426–3445.

[6] Fidalgo, E., Alegre, E., Gonzalez-Castro, V. & Fernández-Robles, L., 2018. "Boosting image classification through semantic attention filtering strategies." Pattern Recognition Letters 112, 176–183.

[7] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F., 2023. "Trocr: Transformer-based optical character recognition with pre-trained models." In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. pp. 13094–13102.

[8] Mishra, S., Soni, D., 2022. "Sms phishing dataset for machine learning and pattern recognition." In: International Conference on Soft Computing and Pattern Recognition. Springer, pp. 597–604

[9] Chollet, F., 2017. "Xception: Deep learning with depthwise separable convolutions." In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258.

[10] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell & T., Xie, S., 2022. "A convnet for the 2020s." In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986.

[11] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G. & Gelly, S., 2020. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale."

[12] Smith, R., 2007. "An overview of the tesseract ocr engine." In: ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition. IEEE Computer Society, Washington, DC, USA, pp. 629–633.

[13] Wang, W., Xie, E., Li, X., Liu, X., Liang, D., Yang, Z., Lu, T. & Shen, C., 2021. Pan++: "Towards efficient and accurate end-to-end spotting of arbitrarilyshaped text." IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (9), 5349–5367.

A review of SpamClus: An agglomerative clustering algorithm for spam email campaigns detection

Alicia Martínez-Mendoza*† amartm@unileon.es
Laura Fernández-Robles*† l.fernandez@unileon.es
Enrique Alegre*† enrique.alegre@unileon.es
Eduardo Fidalgo*† eduardo.fidalgo@unileon.es
Encina Díez-del-Río* mdiezr@unileon.es
*Universidad de León, Campus Vegazana, 24007 León, Spain
†Researcher at INCIBE

***Abstract*—Spam email campaigns can pose a significant cybersecurity threat by enabling cybercriminals to distribute scams and malware at scale, resulting in financial loss and device compromise. Enabling earlier detection and improved response to spam-based threats may be a priority for Computer Emergency Response Teams technicians. This summary presented SpamClus, an iterative algorithm based on unsupervised learning and agglomerative clustering, capable of identifying groups of emails that form coherent spam campaigns. SpamClus combines BERT embeddings to capture semantic relationships among emails with effective thresholds in the minimum number of samples and cosine similarity to determine meaningful clusters. SpamClus has been evaluated on real data from INCIBE, identifying 19 spam campaigns, with 3,048 emails from 6,702 emails over three consecutive days.**

***Index Terms*—spam email detection, agglomerative clustering, cybersecurity**

Type of contribution: *Already published [1]*

I. Introduction

Spam is a significant cybersecurity issue due to its low cost and anonymity, making it a preferred tool for cybercriminals, often leading to financial losses [2]. Companies and platforms like Gmail use awareness campaigns and filters to combat spam, but spammers employ techniques like hidden text, images, or HTML tags to bypass these measures, complicating detection [3]. Spam detection has been approached as both binary classification (spam vs. non-spam) [4] and multi-classification, clustering emails by campaign topics [5], [6]. Methods like URL-based clustering [7] and the Mean Shift algorithm [8] have been used to group emails based on sender, subject, body, and URL.

The paper introduced SpamClus, a novel approach that combines agglomerative clustering, BERT embeddings, and Optical Character Recognition (OCR) preprocessing to detect spam campaigns more effectively. Unlike previous methods that relied on DBSCAN [8] or topic modeling [5], agglomerative clustering groups emails based on similarity, while BERT embeddings capture contextual relationships between emails, replacing traditional features like sender or subject. Additionally, OCR preprocessing removes hidden text, ensuring only visible content is analyzed.

The goal is to identify spam campaigns, not just detect or classify spam by topic. This approach is particularly useful for Computer Emergency Response Teams (CERTs), enabling early detection to mitigate attacks and protect users. By integrating these techniques, SpamClus addresses previous limitations and offers a robust solution for combating spam campaigns.

II. Methodology

A. Datasets and pre-processing

We used a dataset of $4,829$ spam emails provided by INCIBE, which included potential spam campaigns. These emails, collected in 2021, contained both visible and hidden text, with the latter added by spammers to introduce noise and reduce the effectiveness of spam filters [3]. To extract only the visible text, we applied OCR using the Python library Pytesseract. Additionally, we removed special characters, HTML and CSS tags, and replaced personalized information (such as email addresses and URLs) with tokens to minimize differences between emails belonging to the same campaign but without modifying the context.

B. Spamclus

SpamClus evaluates thresholds iteratively, forming clusters to identify potential spam campaigns based on two criteria: a minimum number of samples (`min_samples`) and a minimum average cosine similarity (`min_similarity`) between emails within the cluster. Formally, a cluster C_i is classified as a spam campaign if:

$$(|C_i| \geq \text{min_samples}) \wedge \left(\frac{1}{|C_i|(|C_i|-1)} \sum_{e_j, e_k \in C_i, j \neq k} \text{cos_sim}(e_j, e_k) \geq \text{min_similarity} \right)$$

where $|C_i|$ is the number of emails in cluster C_i, and min_samples and min_similarity are predefined threshold parameters.

- *min_samples*: Minimum number of samples required to consider a cluster as a spam campaign. It is set to 50 based on recommendations from INCIBE experts.
- *min_similarity*: Minimum cosine similarity value required to classify a cluster as a spam campaign. It is set to 0.8 based on tests across different date ranges.

The clustering process begins with the computation of an initial threshold, determined by the Euclidean distance between a randomly selected subset of 300 email samples. Subsequently, the email content undergoes preprocessing, followed by the extraction of dense vector representations (embeddings) using a multilingual BERT model, which handles both English and Spanish emails. Agglomerative clustering groups similar emails iteratively, assigning each to a cluster. The algorithm then evaluates each cluster by

computing the number of samples and the average pairwise cosine similarity within the cluster. Clusters that satisfy predefined criteria—meeting or exceeding a minimum number of samples and similarity threshold—are classified as spam campaigns, stored in a new dataset, and removed from the original dataset. The iterative process continues until either the similarity threshold becomes non-positive or the dataset contains insufficient samples to form meaningful clusters.

III. Results and discussion

To determine the value for *min_similarity*, tests were carried out with different ranges of days, as shown in Tables I and II. The performance of the algorithm was evaluated by varying the value of *min_similarity*, with the objective of grouping spam campaigns while avoiding clusters with mixed topics.

TABLE I
SPAM CLUSTERS WITH DIFFERENT MINIMUM COSINE SIMILARITY (FROM 11/12/2021 TO 13/12/2021)

Min similarity	# Non-Spam Samples	# Spam Samples	# Spam Clusters	Mixed clusters?
0.9	3455	3247	16	No
0.8	**3654**	**3048**	**19**	**No**
0.7	4791	1911	22	Yes
0.6	5093	1609	24	Yes
0.5	5093	1609	24	Yes

TABLE II
SPAM CLUSTERS WITH DIFFERENT MINIMUM COSINE SIMILARITY FOR ONE DAY (11/12/2021)

Min similarity	# Non-Spam Samples	# Spam Samples	# Spam Clusters	Mixed clusters?
0.9	261	1208	4	No
0.8	423	986	8	No
0.7	599	870	8	No
0.6	**599**	**870**	**8**	**No**
0.5	823	646	9	Yes

A *min_similarity* value of 0.8 was found optimal for preventing topic mixing within clusters, particularly when analyzing email data over a three-day period, while a lower threshold of 0.6 worked best for single-day clustering. The algorithm used 0.8 for consistency, ensuring cluster cohesion and reliable spam campaign identification over longer periods. SpamClus was tested on INCIBE's spam campaign dataset. The analysis revealed that, within a three day spam email set containing 6072 emails, the algorithm successfully identified 19 distinct clusters, encompassing a total of 3048 emails. For a one-day dataset of 1469 emails, SpamClus detected 8 clusters, grouping a total of 870 emails. These results demonstrate the algorithm's effectiveness in clustering spam emails over different timeframes, highlighting its capability to identify structured spam campaigns with high accuracy. A visual examination of the campaigns revealed that the emails from the same cluster share the same structure. The SpamClus algorithm proved to be better than a fixed-threshold approach. When we used a high threshold, the smaller campaigns were grouped under the same cluster, resulting in campaigns with mixed topics. The algorithm created small clusters with very high cosine similarity for a low threshold but with insufficient samples to reach the *min_samples* value. To address this issue, the SpamClus algorithm implements an iterative approach that dynamically adjusts the threshold. Concerning the input parameters, a low value for *min_samples* could generate small clusters without representativeness, while a high value could omit legitimate campaigns. For the *min_similarity*, a low value could increase the risk of mixing topics in clusters, while a high value could reduce the campaign detection rate.

IV. Conclusion

The paper introduced SpamClus, an algorithm for detecting spam campaigns using agglomerative clustering, eliminating fixed thresholds, and establishing criteria based on the number of samples and email similarity. An optimal similarity value of 0.8 was identified, allowing the detection of multiple spam clusters over different periods.

The approach is highly relevant to cybersecurity, as it enhances spam detection using clustering techniques, thereby strengthening email security and reducing the risk of cyber threats. By identifying spam email campaigns, SpamClus can alert targeted individuals and organizations, reducing victims. Future improvements include automating similarity calculations, better removing hidden text, and enhancing detection of advanced spam techniques.

Acknowledgements

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation), thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

References

[1] D. Díaz, W. Al-Nabki, L. Fernández-Robles, E. Alegre, E. Fidalgo, and A. Martínez-Mendoza, "SpamClus: An agglomerative clustering algorithm for spam email campaigns detection," in *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2024)*, Lancaster, United Kingdom, 2024. [Online]. Available: https://nlpaics-2024.org/programme-2/

[2] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "Spam email classification based on cybersecurity potential risk using natural language processing," *Knowledge-Based Systems*, vol. 310, p. 112939, 2025.

[3] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: analysis of spammer strategies and the dataset shift problem," *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1145–1173, 2023. [Online]. Available: https://doi.org/10.1007/s10462-022-10195-4

[4] A. Karim, S. Azam, B. Shanmugam, and K. Kannoorpatti, "An unsupervised approach for content-based clustering of emails into spam and ham through multiangular feature formulation," *IEEE Access*, vol. 9, pp. 135 186–135 209, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3116128

[5] A. Ligthart, C. Catal, and B. Tekinerdogan, "Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification," *Applied Soft Computing*, vol. 101, p. 107023, 2021. [Online]. Available: https://doi.org/10.1016/j.asoc.2020.107023

[6] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach," *Applied Soft Computing*, vol. 139, p. 110226, 2023.

[7] F. Li and M.-H. Hsieh, "An empirical study of clustering behavior of spammers and group-based anti-spam strategies," in *CEAS 2006 - The Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California, USA*, vol. 2006, 2006, pp. 21–28. [Online]. Available: https://ceas.cc/2006/1.pdf

[8] K. Althobaiti, K. Vaniea, M. K. Wolters, and N. Alsufyani, "Using clustering algorithms to automatically identify phishing campaigns," *IEEE Access*, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3310810

Comparative analysis of Natural Language Processing methods for Authorship Attribution of texts of different lengths

Alicia Gómez-Pascual*, Víctor González-Castro*, Alicia Martínez-Mendoza*, Christopher Gaul*
{alicia.gomez, victor.gonzalez, amartm, cgau}@unileon.es
*Universidad de León. Campus de Vegazana s/n, 24007 León, Spain

***Abstract*—Authorship Attribution is a critical task in identifying the authors of texts in different contexts, from literary studies to cybersecurity. This paper compares two state-of-the-art models, i.e., Diff-Vectors and Microtext using Capsule Networks, as well as some traditional machine learning methods and Transformers, for this task. We have assessed these methods on three datasets with short, medium and large texts, along with the effects of preprocessing the text. The results obtained indicate that while Diff-Vectors and traditional models such as Logistic Regression and Shallow Neural Networks perform well on longer texts, whereas DistilBERT excels with short and medium length texts. Additionally, simple models like Logistic Regression with Bag of Words demonstrated competitive performance, challenging the supremacy of more complex approaches.**

***Index Terms*—Authorship Attribution, Machine Learning, Deep Learning, Natural Language Processing**

Type of contribution: *Original research in progress*

I. Introduction

Authorship Attribution (AA) is the task of identifying the author of a given text from a group of potential candidates. Historically, AA has been associated with philology, aiding scholars to uncover the authorship of anonymous literary works or resolve disputes related to historical documents and legal cases. However, the rapid proliferation of digital communication and social media platforms has expanded its relevance into the realm of cybersecurity and crime prevention. Today, cybercriminals exploit digital anonymity to carry out malicious activities such as the dissemination of fake news, phishing attacks, spam campaigns, cyberbullying, online grooming, or hate speech propagation. In this complex landscape, AA becomes a powerful forensic tool, helping law enforcement agencies to find evidences to build legal cases by linking textual evidence to individuals behind cybercrimes [1]. Given the huge volume and complexity of digital data generated daily, a manual analysis of textual evidence would be impractical; thus, automated AA techniques leveraging machine learning and Natural Language Processing (NLP) have become essential. Recent advancements in these computational methods have significantly increased their accuracy and efficiency, making them indispensable in modern cybersecurity investigations [2].

In AA tasks, there are many possible approaches. Some works in the literature use traditional methods like n-grams [3], while others prefer deep learning techniques, such as Convolutional Neural Networks [4] and Transformers, as seen in *Figuratively Speaking* [5].

In this paper, we present a *work in progress* about AA. Specifically, we have reviewed the literature in this field, and compared some state-of-the-art methods with other more traditional NLP ones on datasets for different text sizes. While the selected datasets are not drawn specifically from cybersecurity contexts, they reflect the challenges that AA techniques face in such scenarios. The selected datasets contain short, medium and long texts, which mirror the length of the texts that are found in different cybersecurity problems or cibercrimes, such as smishing or grooming, fake reviews or anonymous manifestos or phishing emails, respectively.

The rest of the paper is structured as follows: In Section II we review the state-of-the-art in AA. The methodology of the experiments we have carried out is explained in Section III, along with some details about the specific methods we have assessed and the datasets we have used. In Section IV we show the results of those experiments, which we discuss in Section V. Finally, the conclusions of this work are presented in Section VI.

II. Literature review

The goal of AA is to determine the person that most likely wrote a given text. There is a similar task called Authorship Verification (AV), which focuses on verifying whether the text was written by a specific author or not. In the International Symposium on Intelligent Data Analysis, the model LG4AV [6] was presented to address the AV challenge using pre-trained Transformers and graph neural networks. For the same purpose, Almutairi et al. developed BiBERT-AV [7], which used Siamese methods, pre-trained Transformer models and Bi-LSTM layers.

To carry out the AA task, some models use the pre-trained Transformer BERT [8], like BertAA [9], or BERTPDS [10]. Although BERTPDS was developed to detect plagiarism, it uses algorithms for AA to support the decision of whether it is plagiarism or not

Other works focus on linguistic features. Katsios et al. [5] developed a multi-task figurative language model that detects multiple figurative language features such as personification or metonymy. Eder and Górski [3] tried to improve the results of AA by means of techniques such as grammatical features or lemmatization, finding out that, in the corpus of Polish novels they assessed, these features worked worse than lexical features. In addition, Zhang et al. proposed a Syntax Encoding method [11], which constructed an embedding for each word that encoded its syntax parse tree.

While most models focus on static representations, others have explored different approaches. For instance, Terreau and Velcin developed a new representation model using Brownian Bridges [12]. In addition, Corbara et al. investigated the impact of using diff-vectors [13], which are feature vectors that represent the difference between the existing pairs of objects in the features: function words, word lengths, sentence lengths, punctuation symbols, POS n-grams, word unigrams, and character n-grams.

Because of the usefulness of finding the author of short texts, some models, such as the one developed by Suman et al. [4], specialised in microtexts like tweets. To achieve this, they utilised two architectures: Capsule-Based CNN, which uses a Convolutional layer, and Capsule-Based KNN, which incorporates a kervolutional layer performing the convolution in a nonlinear space. Additionally, the model AA-CNN specialised in predatory text messages [14]. Also, Aouchiche et al. developed three models, i.e., Autoencoder-Adaboost model, Temporal Convolutional Networks (TCN) and another, which combined Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), to determine the authors of tweets [15]. On the other hand, very long texts can be a problem, therefore Kettaf and Yousfate developed an extension of the Linear Discriminant Analysis (FDA) [16].

III. Methodology

In this paper, we compared two state-of-the-art methods, i.e., Diff-Vectors [13] and Microtext using Capsule Networks [4], selected for their recent strong performance in the authorship attribution literature. We also assessed the Transformer Distilbert [17], chosen for its effectiveness as a lightweight alternative to larger Transformers. In addition, we evaluated traditional classification and description methods, which have consistently demonstrated strong performance in a wide range of NLP tasks. This comparison has been carried out using three datasets: the Tweets Dataset [18], the IMDB62 Dataset [19], and the Victorian Era Dataset [20]. In this Section, we present the methods, the datasets and the experiments we carried out.

A. Methods

The method Diff-Vectors [13] uses learners such as Support Vector Machine (SVM), Logistic Regression (LR), and Multinomial Naive Bayes (MNB) to classify the texts using the textual features: function words, word lengths, sentence lengths, punctuation symbols, POS n-grams, word unigrams, and character n-grams. Then, to attribute the authorship, four methods are presented in [13]: PairLRknn (Pknn), PairLR-linear (Plinear), LR and LRbin, although we found that the last two present the same implementation in the published code [21]. The authors of [13] utilized four datasets in their experiments and found that, in some cases, Diff-Vectors improved the AA, but were not suitable for all scenarios.

The authors of Authorship Attribution of Microtext Using Capsule Networks [4] proposed eight new models by combining GloVe, BERT, Unigram and Bigram with two types of neural network architectures: Capsule-Based CNN and Capsule-Based KNN. In the CNN architecture, the data is passed through a convolutional layer, followed by a pooling layer. The output is then sent to a convolutional capsule, and the results are passed to a Dense layer for classification with Softmax. The KNN architecture is similar to the CNN architecture, but the replaces the convolution with a *kervolution*, a kernel + convolution operation, which realizes the convolution in a non-linear space.

B. Datasets

AA can be applied to texts with plenty of variation, e.g., it could be used for short texts in social media, or for longer texts. To assess the models in different situations, we decided to use three datasets that present texts with different lengths and contents.

The Tweets Dataset [18] contains 9,908 tweets written by five celebrities. With this dataset we can study the behavior of the models with short texts written in informal situations.

The IMDB62 Dataset [19] is a well-known dataset for AA containing reviews of films and series carried out by 18 different people. Such reviews contain positive and negative comments, and a wide variety of ratings, allowing for the analysis of the author's writing across a wide range of emotions.

To assess AA in long texts, we have used the Victorian Era Dataset [20]. Each instance has 1000 words from an author's book. There are 45 different authors in the training dataset and 50 in the test dataset. In this paper, we use the 45 authors that appear in the training set.

In our experiments, we have used the original texts of the datasets in the training and test of the models. Additionally, we created two variations of these datasets. The first one involved augmenting the data by splitting the texts into smaller segments, with a maximum of 512 characters each. Since the longest text in the Tweets dataset was 199 characters, this approach was applied only to the IMDB62 and Victorian datasets, but not to the Tweets dataset. The second one consisted of preprocessing the text by removing URLs, emojis, special characters, punctuation marks, and stop words, as well as applying stemming and expanding contractions.

It is known that augmenting the data and preprocessing the text may be useful in NLP tasks. Here, however, we want to re-evaluate the usefulness of augmentation and preprocessing in AA, because it might remove intrinsic characteristics of the author's writing style.

Table I presents more information about the Datasets and how the number of samples, maximum length and mean length are affected by the two different processing techniques mentioned above.

C. Experiments

We divided each dataset into training and testing subsets, using 70% of the data for training and the remaining for testing.

Apart from the models listed in Section III-A, we have also assessed the performance of the pretrained model DistilBERT [17], both in its cased and uncased versions.

In addition, we have evaluated the performance of some traditional Machine Learning models. In particular, we used LR with a regularization strength between 10^{-3} and 10^{3} on a logarithmic scale. We also tested a Naïve Bayes classifier with a Gaussian distribution. Furthermore, we trained a Random Forest classifier, varying the number of estimators between 10

Table I
INFORMATION ABOUT EACH VARIATION OF THE ASSESSED DATASETS, INCLUDING THE TOTAL NUMBER OF SAMPLES, AVERAGE AND COEFFICIENT OF VARIATION OF THE SAMPLES PER CLASS, AND THE MAXIMUM AND AVERAGE LENGTH. THE COEFFICIENT OF VARIATION WAS CALCULATED AS THE RATIO OF THE STANDARD DEVIATION TO THE MEAN, EXPRESSED AS A PERCENTAGE. THE LENGTH OF THE TEXT IN THE LAST TWO COLUMNS IS EXPRESSED IN THE NUMBER OF CHARACTERS.

Dataset	Labels	Dataset variation	Samples	Samples/class (mean)	Samples/class (Coefficient of variation)	Max. Length	Mean Length
Tweets	5	Original	9908	1981.60	1.86%	199	116.41
		Preprocessing	9908	1981.60	1.86%	146	66.19
IMDB62	18	Original	18089	952.05	21.24%	11646	1672.96
		Data Augmentation	71386	3965.89	67.30%	512	394.82
		Preprocessing	17089	952.05	21.24%	6938	950.13
Victorian	45	Original	53678	1192.84	99.35%	5760	4942.97
		Data Augmentation	188969	4199.31	109.56%	512	367.49
		Preprocessing	53678	1192.84	99.35%	4009	2544.36

and 2000. Additionally, we evaluated a Shallow Neural Network, with a hidden layer size ranging from 8 to 512 neurons and ReLU activation function, followed by an output layer with Softmax activation function, and the Adam optimizer. Lastly, we evaluated a Support Vector Machine with linear, polynomial, radial basis function, and sigmoid kernels.

The configuration of the models of the state-of-the-art was explained in Section III-A, and the datasets we used in the assessment were mentioned in Section III-B.

IV. RESULTS

In this Section, we show the results of the different models evaluated on the datasets explained in Section III. Firstly, we assess the performance of the different combinations of method-learner in Diff-Vectors. Secondly, we compare the best learner-method combination of Diff-Vectors with the Attribution of Microtext Using Capsule Networks model and DistilBERT using different preprocessing techniques on the texts. Next, we use Diff-Vectors and LR to evaluate the performance of different embeddings. Finally, we compare traditional machine learning models with the Bag of Words descriptor.

First, since Diff-Vectors [13] may use three possible learners and four possible methods, as explained in Section III-A, we compare the performance of the different combinations. The results of this comparison with the original datasets (i.e., without splitting or preprocessing of the texts) are presented in Table II, except for those of the learner MNB, as they are significantly inferior to those of the other two. The results of the LR and LRbin methods are presented together under the name LR, as they share the same code implementation. The results for methods PairLRknn (Pknn) and PairLRlinear (Plinear) in the Victorian dataset could not be obtained due to limitations on computational resources.

Next, we compare the results obtained with the best learner-method combination of Diff-Vectors (i.e., LR learner and LR method), with the models proposed in Authorship Attribution of Microtext Using Capsule Networks [4], considering six out of the eight methods from the original study due to computational limitations, as well as the Transformer model DistilBERT. All models are evaluated using different processing techniques on the datasets mentioned in Section III-B.

In Table III we show the accuracies obtained using the variations of the different datasets. Data Augmentation was not applied to the Tweets dataset, as it was explained in Section III-B. We can observe that data augmentation in IMDB62 and Victorian datasets worsens the results compared to the original dataset in all cases, except those in the Victorian Dataset where the performance on the original dataset is below 40%. Also, preprocessing the text in the Tweets and IMDB62 datasets worsens the accuracy, but in the Victorian Dataset only worsens the results when using DistilBERT, and improves the results with DiffVectors, CNN and KNN.

Table II
DIFFERENT LEARNERS AND METHODS IN DIFF-VECTORS [13]. THE PERFORMANCE HAS BEEN MEASURED IN ACCURACY (ACC) AND F1-MACRO SCORE (F1), BOTH IN %

Learner	Method	Tweets		IMDB62		Victorian	
		Acc	F1	Acc	F1	Acc	F1
SVM	Pknn	63.30	62.92	80.00	77.24	—	—
SVM	Plinear	85.81	85.67	95.26	94.60	—	—
SVM	LR	82.37	82.26	93.49	92.99	89.75	88.11
LR	Pknn	73.43	73.28	86.51	85.90	—	—
LR	Plinear	89.14	89.11	**96.53**	**96.52**	—	—
LR	LR	**90.51**	**90.44**	91.32	90.17	**93.85**	**91.14**

We present the results of another experiment in which we assess different embeddings using the Diff-Vectors method. In addition to the default embeddings used in the available code [21], TF-IDF, we have also assessed BOW and word2vec. Table IV shows that using BOW, the results are improved in the three datasets. Given the good performance that BOW embeddings have shown with Diff-Vectors, we have evaluated the performance of such embeddings with the LR classifier. As it is shown in Table V, the results are comparable.

Noticing the good results yielded by a traditional machine learning model such as LR with BOW, we have evaluated more machine learning models with the BOW descriptor. In Table VI, it is shown that the Shallow Neural Network and LR obtain the best results.

In Table VII we compare the combinations that obtained the best results in Tables II and IV in the Diff-Vectors model, along with the top-performing machine learning models. Additionally, we include the Microtext Using Capsule Networks methods and the DistilBERT Transformers tested in our study. We can see that the model that obtains the best accuracy with Tweets and IMDB62 datasets is the Transformer DistilBERT [17] without processing the text. Conversely, the Victorian dataset yields better results using machine learning models or Diff-Vectors [13]; it has a similar performance, either preprocessing the text or not, and using Logistic Regression

Table III
ACCURACIES (IN %) OF DIFFERENT MODELS WITH THE ASSESSED DATASETS VARIATIONS. THE BEST-PERFORMING METHOD FOR EACH DATASET VARIATION IS SHOWN IN BOLD.

Method	Tweets		IMDB62			Victorian		
	Original	Preprocessing	Original	Data Augmentation	Preprocessing	Original	Data Augmentation	Preprocessing
Diff-Vectors LR LR [13]	90.51	88.40	91.32	85.25	**92.28**	**93.85**	64.93	**96.44**
CNN GloVe [4]	71.59	65.35	67.06	61.19	59.53	42.05	37.64	43.67
CNN Unigram [4]	90.11	85.93	69.31	60.39	55.65	22.03	37.60	27.93
CNN Bigram [4]	80.63	77.93	64.80	61.82	53.52	22.06	39.62	27.12
KNN GloVe [4]	81.03	75.41	90.61	75.71	82.23	75.27	57.19	82.83
KNN Unigram [4]	88.65	82.37	62.17	53.13	48.81	22.84	35.80	28.89
KNN Bigram [4]	77.50	77.33	61.12	58.35	51.26	23.68	36.39	28.29
DistilBERT uncased [17]	**94.72**	**89.77**	**98.42**	**92.66**	91.13	92.49	**72.03**	91.65
DistilBERT cased [17]	92.87	89.74	98.11	92.23	90.99	91.05	**72.03**	90.54

Table IV
ACCURACY (ACC) AND F1-SCORE (F1), IN %, OF DIFFERENT EMBEDDINGS USED WITH DIFF-VECTORS [13] USING LEARNER LR AND METHOD LR.

Embedding	Tweets		IMDB62		Victorian	
	Acc	F1	Acc	F1	Acc	F1
TF-IDF	**90.51**	**90.44**	91.32	90.17	93.85	91.14
BOW	89.14	89.13	**95.24**	**94.99**	**97.00**	**96.35**
Word2vec	68.21	67.84	90.68	90.04	80.04	76.13

Table V
DIFFERENT EMBEDDINGS IN LOGISTIC REGRESSION

Embedding	Tweets		IMDB62		Victorian	
	Acc	F1	Acc	F1	Acc	F1
TFIDF	**90.21**	**90.16**	92.71	91.50	92.98	90.14
BOW	89.14	89.13	**95.24**	**94.98**	**96.91**	**96.19**
Word2vec	68.45	68.10	90.70	89.99	80.13	76.20

or Diff-Vectors.

V. DISCUSSION

In our experiments, we found that the performance of the models varied according to the text size. For this reason, we selected three datasets with different text lengths, having short texts in the Tweets Dataset, medium-length texts in the IMDB62 Dataset and long texts in the Victorian Dataset. We observed that models specialized in short text obtain competitive performance in short text datasets, but are unable to obtain similar results in more lengthy texts. For example, M-I-C (CNN) obtained 90.11% accuracy in the Tweets dataset, but less than 30% with the larger texts. The Transformer-based model DistilBERT worked the best with medium datasets, obtaining the highest result in the IMDB62 (98.42%), but also showed a good performance in short and large datasets achieving 94.72% and 92.49% accuracy, respectively. Moreover, we can see that Diff-Vectors, Logistic Regression and Shallow Neural Network obtained the best results in the dataset with long texts, between 96% and 97% accuracy in the Victorian Dataset, while in the shorter dataset they obtained accuracy values between 89% and 91%.

Another observation is about processing the text. We did not obtain any improvement by splitting the text into shorter text and obtaining more samples to train the model. With respect to the preprocessing explained in Section III-B, the performance of the models decreased in the Tweets and IMDB62 datasets. However, for the Victorian dataset, some models showed nearly identical results, whether the text was preprocessed or left in its original form.

Table VI
ACCURACY (ACC) AND F1-SCORE (F1) OF TRADITIONAL MACHINE LEARNING CLASSIFIERS WITH THE BOW DESCRIPTOR

Model	Tweets		IMDB62		Victorian	
	Acc	F1	Acc	F1	Acc	F1
LR	89.14	89.13	95.24	94.98	**96.91**	**96.19**
Naive Bayes	75.14	75.00	28.96	24.11	27.23	11.94
Random forest	76.52	76.81	76.03	70.97	42.39	22.22
Shallow NN	**90.18**	**90.11**	**96.86**	**95.92**	96.09	94.83
SVM	78.78	78.77	80.89	79.33	76.94	69.77

It is remarkable that traditional classification and description methods have demonstrated good performance. For example, changing the original Diff-Vectors TF-IDF embedding by a simple one such as BOW improves its performance by around 4% in the IMDB62 and Victorian datasets. Moreover, we observed that LR achieved very similar results to Diff-Vectors and comparable to those of the Transformer-based DistilBERT. In particular, LR obtained better results than DistilBERT in longer texts (96.91% against 92.49% accuracy on the Victorian dataset), but worse results in shorter texts (86.07% and 89.77% accuracy, respectively).

VI. CONCLUSION

We have reviewed the state-of-the-art in Authorship Attribution (AA), comparing the performance of the state-of-the-art models Diff-Vectors and Microtext Using Capsule Networks. We have also assessed other methods, i.e. traditional machine learning with TF-IDF and BOW, and DistilBERT. We have compared all these methods on the datasets Tweets, IMDB62, and Victorian, which have texts with different lengths (short, medium-length and long, respectively). This comparison shows that some models are more specialized in short texts, while others perform well with longer text sizes.

We also observed that, in general, preprocessing the text does not improve the performance in AA. In the case of the Tweets and IMDB62 datasets, it deteriorated the performance. In contrast, in the Victorian dataset, it had little or no effect on the results.

An interesting finding is the effectiveness of simple methods, such as BOW and LR, in this task, obtaining results that can compete with more complex methods like Diff-Vectors and DistilBERT.

Table VII
ACCURACY (ACC) AND F1-SCORE (F1) OF THE MAIN METHODS USING DIFFERENT TEXT PROCESSING TECHNIQUES

Model	Embedding	Tweets				IMDB62				Victorian			
		Default		Preprocess		Default		Preprocess		Default		Preprocess	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
D-V Plinear LR [13]	Default	89.14	89.11	81.97	81.91	96.53	89.46	89.16	89.46	—	—	—	—
D-V LR LR [13]	Default	90.51	90.44	88.40	88.35	91.32	90.17	92.28	91.74	93.85	91.14	96.44	95.44
D-V LR LR [13]	BOW	89.14	89.13	86.07	86.06	95.24	94.99	93.99	93.24	**97.00**	**96.35**	96.96	**96.21**
Logistic Regression	BOW	89.14	89.13	86.07	86.06	95.24	94.98	**94.27**	**93.83**	96.91	96.19	**96.99**	96.20
Shallow Neural Network	BOW	90.18	90.11	84.33	84.24	96.86	95.92	81.08	74.38	96.09	94.83	74.90	57.24
CNN (M-I-A) [4]	GloVe	71.59	71.10	65.35	65.33	67.06	63.31	59.53	57.32	42.05	25.48	43.67	29.61
CNN (M-I-C) [4]	Unigram	90.11	90.11	85.93	84.98	69.31	68.14	55.65	54.59	22.03	13.75	27.93	18.38
CNN (M-I-D) [4]	Bigram	80.63	80.60	77.93	77.75	64.80	63.60	53.52	51.15	22.06	11.96	27.12	14.13
KNN (M-II-A) [4]	GloVe	81.03	81.12	75.41	75.26	90.61	90.39	82.23	81.16	75.27	68.75	82.83	77,69
KNN (M-II-C) [4]	Unigram	88.65	88.67	82.37	82.45	62.17	60.81	48.81	48.45	22.84	14.97	28.89	20.33
KNN (M-II-D) [4]	Bigram	77.50	77.26	77.33	77.16	61.12	57.90	51.26	48.73	23.68	9.59	28.29	13.69
DistilBERT uncased [17]	None	**94.72**	**94.71**	**89.77**	**89.76**	**98.42**	**98.42**	91.13	91.24	92.49	92.48	91.65	91.64
DistilBERT cased [17]	None	92.87	92.88	89.74	89.75	98.11	98.11	90.99	91.06	91.05	91.04	90.54	90.53

The task of AA has the potential to assist in the detection of predators in online conversations, enhancing the security of the online environment. Moreover, it could offer significant value to Law Enforcement Agencies by providing a tool to automatically analyse text conversations in ongoing investigations.

Due to computational resource limitations, we only performed a train-test split for this study. Therefore, in future works, we plan to implement cross-validation for a more robust evaluation. Additionally, we will perform a more in-depth analysis of the performance of AA on shorter texts, including chat messages, specifically those involving criminal behavior such as grooming. Additionally, we will evaluate other Transformer-based models for the task of AA. Lastly, given the efficacy of machine learning models, we will combine the embeddings generated by Transformers with traditional classification techniques.

ACKNOWLEDGEMENTS

This work has been funded by the Recovery, Transformation, and Resilience Plan, funded by the European Union (Next Generation) thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

REFERENCES

[1] C. E. Chaski, "Who's at the keyboard? authorship attribution in digital evidence investigations," *Int. J. Digit. EVid.*, vol. 4, 2005.

[2] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 5–33, 2017.

[3] M. Eder and R. L. Górski, "Stylistic Fingerprints, POS-tags and Inflected Languages: A Case Study in Polish," *Journal of Quantitative Linguistics*, 6 2022.

[4] C. Suman, A. Raj, S. Saha, and P. Bhattacharyya, "Authorship Attribution of Microtext Using Capsule Networks," *IEEE Transactions on Computational Social Systems*, vol. 9, pp. 1038–1047, 2022.

[5] G. Katsios, N. Sa, and T. Strzalkowski, "Figuratively Speaking: Authorship Attribution via Multi-Task Figurative Language Modeling," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13 240–13 255.

[6] G. S. Maximilian and Stumme, "LG4AV: Combining Language Models and Graph Neural Networks for Author Verification," in *Advances in Intelligent Data Analysis XX*, Elisa, H. E. B. Tassadit, and Fromont, Eds. Springer International Publishing, 2022, pp. 315–326.

[7] BooJoong, N. A. H. A. Amirah, and Kang, "BiBERT-AV: Enhancing Authorship Verification Through Siamese Networks with Pre-trained BERT and Bi-LSTM," in *Ubiquitous Security*, Haozhe, M. Geyong, G. Nektarios, M. W. W. Guojun, and Wang, Eds. Springer Nature Singapore, 2024, pp. 17–30.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[9] M. Fabien, E. Villatoro-Tello, P. Motlicek, and S. Parida, "BertAA : BERT fine-tuning for Authorship Attribution," in *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, P. Bhattacharyya, D. M. Sharma, and R. Sangal, Eds. NLP Association of India (NLPAI), 12 2020, pp. 127–137.

[10] A. Bohra and N. C. Barwar, "A Deep Learning Approach for Plagiarism Detection System Using BERT," in *Congress on Intelligent Systems*, M. Saraswat, H. Sharma, K. Balachandran, J. H. Kim, and J. C. Bansal, Eds. Singapore: Springer Nature Singapore, 2022, pp. 163–174.

[11] R. Zhang, Z. Hu, H. Guo, and Y. Mao, "Syntax Encoding with Application in Authorship Attribution," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2742–2753.

[12] E. Terreau and J. Velcin, "Building Brownian Bridges to Learn Dynamic Author Representations from Texts," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14641 LNCS. Springer Science and Business Media Deutschland GmbH, 2024, pp. 230–241.

[13] S. Corbara, A. Moreo, and F. Sebastiani, "Same or Different? Diff-Vectors for Authorship Analysis," *ACM Trans. Knowl. Discov. Data*, vol. 18, 9 2023.

[14] K. Misra, H. Devarapalli, T. R. Ringenberg, and J. T. Rayz, "Authorship Analysis of Online Predatory Conversations using Character Level Convolution Neural Networks," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 623–628.

[15] R. I. A. Aouchiche, F. Boumahdi, M. A. Remmide, and A. Madani, "Authorship attribution in twitter: a comparative study of machine learning and deep learning approaches," *International Journal of Information Technology*, vol. 16, pp. 3303–3310, 2024.

[16] C. Kettaf and A. Yousfate, "Authorship Attribution by Functional Discriminant Analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11989 LNCS. Springer, 2020, pp. 438–449.

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[18] M. A. Haque, "Tweets data for authorship attribution modelling," 2021.

[19] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship Attribution with Topic Models," *Computational Linguistics*, vol. 40, pp. 269–310, 2014.

[20] A. Gungor, "Victorian Era Authorship Attribution," UCI Machine Learning Repository, 2018, DOI:10.24432/C5SW4H.

[21] S. Corbara, A. Moreo, and F. Sebastiani, "Diff-Vectors for Authorship Analysis," 2023.

A Comparative Study of Natural Language Processing Models for Detecting Obfuscated Text

Rubén Valcarce-Álvarez*, Francisco Jáñez-Martino*, Rocío Alaiz-Rodríguez*, Víctor González-Castro*
{rvalca, francisco.janez, rocio.alaiz, victor.gonzalez}@unileon.es
*Universidad de León. Campus de Vegazana s/n, 24007 León, Spain

***Abstract*—Cybersecurity challenges, including scams, drug sales, the distribution of child sexual abuse material, fake news, and hate speech on both the surface and deep web, have significantly increased over the past decade. Users who post such content often employ strategies to evade detection by automated filters. Among these tactics, text obfuscation plays an essential role in deceiving detection systems. This approach involves modifying words to make them more difficult for automated systems to interpret while remaining sufficiently readable for human users. In this work, we aim at spotting obfuscated words and the employed obfuscation technique, such as leetspeak, word inversion, punctuation changes, or a mix of the previous ones. We benchmark Named Entity Recognition (NER) using Transformer models on the XX-NER-WordCamouflage dataset. Our experiments evaluate these models by comparing their precision, recall, F1 scores, and accuracy, both overall and for each individual class.**

***Index Terms*—Deep Learning, Named Entity Recognition (NER), Natural Language Processing (NLP), Text Obfuscation.**

Type of contribution: *Original research in progress*

I. Introduction

The applications of text obfuscation span across multiple domains. Privacy-focused applications may use obfuscation to anonymize sensitive textual data, particularly in medical and legal records [1], [2]. Cybercriminals operating on the dark web rely on anonymous Tor traffic to hide illicit activities [3], but they may also use text obfuscation in their online activities. Obfuscation techniques are also used to bypass content moderation systems in social media and online forums [4], [5]. Research has shown that automated moderation tools struggle to detect obfuscated hate speech, as seen in studies on profanity obfuscation in academic publications [6].

The exponential rise in cybersecurity threats such as scams, hate speech, and fake news demands the development of effective detection systems, particularly in identifying obfuscated text used to deceive automated filters.

This study addresses the effectiveness of Natural Language Processing (NLP) models in identifying obfuscated text, providing insights into their performance in detecting camouflaged words employed in deceptive online content. It is a work in progress that aims to compare the performance of various NLP models in detecting obfuscated words manipulated through techniques like leetspeak, inversion, and punctuation changes. Specifically, this research evaluates Named Entity Recognition (NER) using BERT-based models on the XX-NER-WordCamouflage dataset [5]. This work provides insights into the effectiveness of Transformer-based models in spotting obfuscated words commonly found in online threats like scams and hate speech.

The rest of the paper is organized as follows. Section II presents a literature review of text obfuscation. The methodology followed is presented in Section III. Section IV describes the dataset and the models evaluated for detecting obfuscated text, along with the results and discussion. Section V summarizes the main contributions of our work and outlines directions for future research.

II. Background

Text obfuscation has been studied recently as a critical component of Natural Language Processing applications [5]. Several works have explored different techniques to modify text while maintaining its readability, primarily for privacy preservation, adversarial text generation, and security purposes [7], [6], [3].

One of the earliest methods involves character-level transformations such as leetspeak and character substitution, which have been widely used to evade detection in online platforms [8], [9]. These techniques replace certain letters with visually similar symbols or numbers, making automated content moderation more challenging (e.g., 'cocaine' becomes 'ç0ç41n3'). More advanced techniques, such as punctuation insertion and syllable inversion, have been introduced to create increasingly complex obfuscations [5], [6], [10] (e.g., 'cocaine' becomes 'c o c a i n e', 'eniacoc' or 'cociane').

At the word level, researchers have explored synonym substitution and masked language modeling [7], [11] to create obfuscated versions of text while preserving semantic integrity. Studies on euphemistic phrase detection further highlight the challenge of identifying obfuscated content in sensitive domains, such as commercial sex advertisements and hate speech detection [6], [12].

Obfuscation at sentence level has also been investigated, with paraphrasing models playing a significant role in generating alternative text structures to evade detection. The ParaDetox dataset has shown effectiveness in improving detection [13].

Stylometric obfuscation has emerged as a method to evade authorship attribution by modifying writing style features such as sentence length, passive-to-active voice transformations, and vocabulary choices [14].

To mitigate the effects of text obfuscation techniques, researchers have developed detection methods that incorporate both rule-based and machine learning approaches. Huertas et al. [5] introduced *pyleetspeak*, a multilingual tool that automatically obfuscates the keywords of a text by means of different techniques, i.e., leetspeak, punctuation and inversion. They used it to create a multilingual dataset to assess their Transformer-based model, *MPNET-ideal*, which achieved an F1-score of 0.8795 across multiple languages, demonstrating the efficacy of Transformers in detecting obfuscated text.

Obfuscation reversal has also been explored. *O-Norm* is a preprocessing tool that normalizes offensive words by reversing obfuscations, achieving an 89.6% F1-score [10]. In [15] an LSTM-based approach is introduced for translating neglected languages, including leetspeak and reverse writing.

In cybersecurity and cybercrime prevention, NLP tasks are very common, e.g., to detect illegal websites or detecting hate speech. Detoxification models such as ParaGeDi and CondBERT by [16] improved text detoxification using the Jigsaw Toxic Comment and ParaNMT datasets. Similarly, [17] employed a BART-based model to generate non-toxic paraphrases, achieving high fluency and content preservation scores. Euphemism detection for censorship evasion was addressed in [18] using a three-stage model. They extracted candidate phrases with AutoPhrase [19], filtered them using word2vec embeddings, and ranked them with SpanBERT. Tested on a drug-related dataset, their model effectively identified euphemisms like "blue dream" for marijuana, enhancing content moderation.

Recent developments in NLP have moved from word embedding models, e.g., Word2Vec and FastText, to Transformer-based models like BERT [20], RoBERTa [21] or DistilBERT [22]. Transformers are now the state of the art for text classification using pre-trained architectures that automate key steps like preprocessing, feature extraction, and classification.

III. Methodology

In this paper, we employ a text classification methodology focused on identifying and classifying obfuscation techniques applied to words within sentences. Our goal is to evaluate several approaches and compare them to the one introduced by Huertas et al. and their LeetSpeaker model [5].

Following a similar approach to Huertas et al. [5], we frame this task as a Named Entity Recognition (NER) classification problem. Specifically, we define four main categories representing the most relevant obfuscation techniques:

- **LEETSPEAK**: Involves transforming words by substituting alphabetic characters with visually or phonetically similar letters, numbers, or symbols (e.g. 'animal' becomes '4N1m4l' or '@n!M@l').
- **INV_CAMO** (Inverse Camouflage): Characterized by reversing the order of characters within a word to obfuscate its meaning (e.g., 'animal' becomes 'aminal' or 'headphones' becomes 'phoneshead').
- **PUNCT_CAMO** (Punctuation Camouflage): Defined by inserting unnecessary punctuation within words, disrupting recognizable patterns (e.g., 'animal' becomes 'a!n!i!m!a!l' or 'headphones' becomes 'hé._a¨Dpho!n?e_s').
- **MIX**: Represents words that combine techniques from both leetspeak and punctuation camouflage.

IV. Experimentation

A. Dataset

We performed the experiments on the **XX_NER_WordCamouflage** Dataset [5]. It is a synthetic multilingual dataset designed to detect word camouflage techniques in over 20 languages. However, although it is a multilingual dataset, we implemented an English language filter to simplify debugging process. It includes data generated from various sources like OPUS News-Commentary, TED2020, and WikiMatrix, and applies camouflaging techniques such as leetspeak, punctuation insertion, and syllable inversion using the pyleetspeak package. The dataset is divided into training, validation, and test sets. The samples and their distribution among these sets have been used as indicated by Huertas et al. [5]. This dataset was chosen because it is the only one in the state of the art that labels words by their obfuscation, which is useful for training models in recognizing camouflaged content.

B. Configuration

The experiments were carried out on a computer with two AMD 7352 processors, each with 24 cores and a clock speed of 2.30GHz; and two NVIDIA Ampere A100 GPUs, each with 40GB of VRAM.

The models selected for the experimental phase are Transformer models, specifically **BERT** [20], **DistilBERT** [22], **MPNET** [23], **RoBERTa** [21] and **XLNet** [24]. We fine-tuned them following the default configuration defined in their corresponding papers.

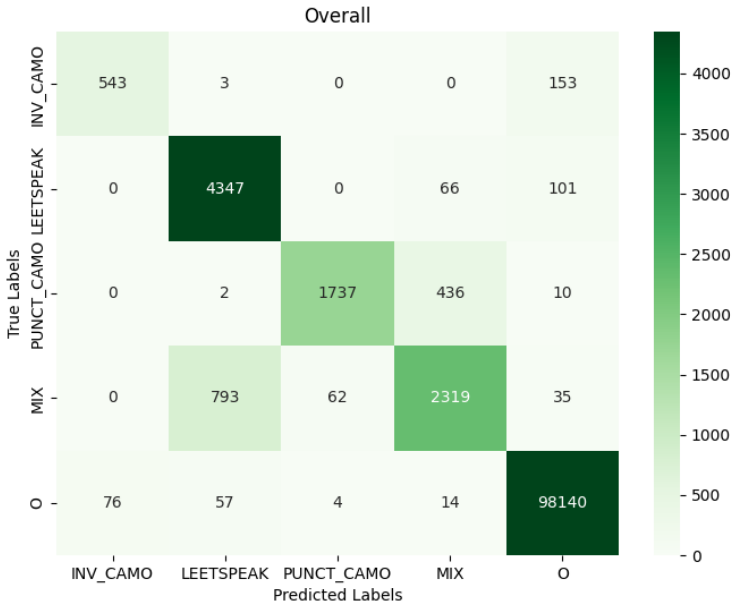
We evaluated these models using the macro, micro and weighted F1-score. The macro F1 score calculates the mean of F1 scores from each class, while the micro F1 score takes into account the confusion matrix across all classes before computing a single F1 score. On the other hand, weighted F1 score considers class size by assigning more weight to classes with more instances. Due to dataset imbalance, macro and weighted F1 scores are preferred: macro F1 treats all classes equally, while weighted F1 prioritizes larger classes. Since camouflage types differ in proportion, the weighted F1 is prioritized. The results are reported both as an overall score and by data set.

C. Results and discussion

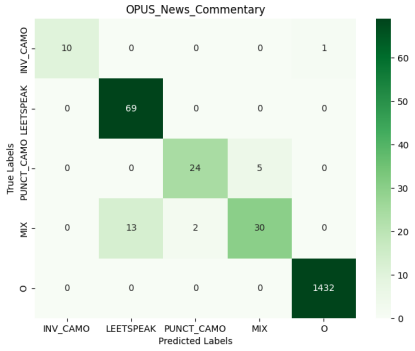
Table I: F1 scores (macro, micro, weighted) for five selected models on each test dataset. The highest value in each row is in bold.

Dataset	F1	BERT	DistilBERT	MPNet	RoBERTa	XLNet
Overall	macro	0.8722	0,8779	0,8487	**0,8783**	0.8587
	micro	0.9815	0.9815	0.9762	**0.9835**	0.9808
	weighted	0.981	0.9811	0.9758	**0.9835**	0.9805
OPUS News Commentary	macro	0.8922	0.8846	0.8536	**0.8977**	0.8897
	micro	0.9855	0.983	0.9792	**0.9868**	0.9849
	weighted	0.9848	0.9825	0.9787	**0.9862**	0.9843
OPUS ParaCrawl	macro	0.8262	**0.8498**	0.8086	0.8485	0.8123
	micro	0.9776	0.9824	0.9739	**0.9842**	0.9791
	weighted	0.9778	0.9824	0.9743	**0.984**	0.9795
TED 2020	macro	**0.8879**	0.8877	0.8742	0.8875	0.8741
	micro	0.983	0.9821	0.9795	**0.9845**	0.9825
	weighted	0.9823	0.9817	0.9789	**0.9841**	0.9821
WikiMatrix	macro	0.8828	**0.8875**	0.8586	0.8794	0.859
	micro	0.9781	0.9786	0.9722	**0.9804**	0.9768
	weighted	0.9774	0.9781	0.9714	**0.9798**	0.9761

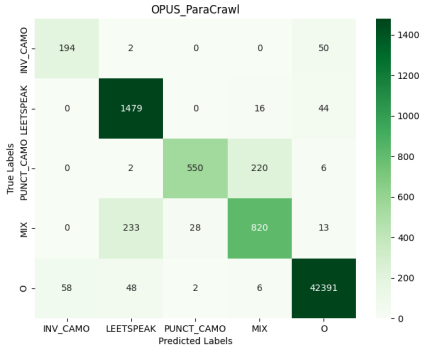
Table I reports the performance results. We can see that RoBERTa consistently outperforms the others experimental models, achieving notably high F1 scores. BERT and DistilBERT closely follows, exhibiting a similar strong performance, in some cases even surpassing it in macro F1 scores. On the other side, MPNet and XLNet, generally demonstrate lower performance across metrics. However, they have also



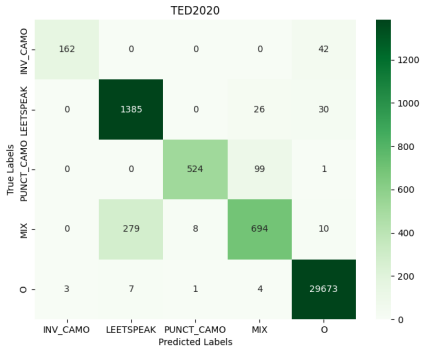
(a) Overall



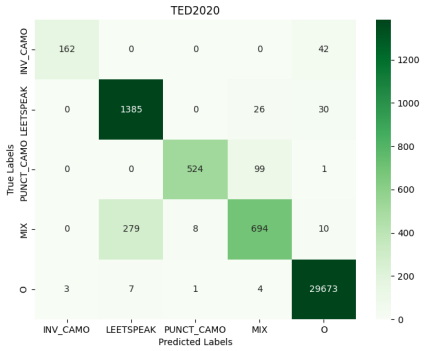
(b) News Commentary



(c) ParaCrawl

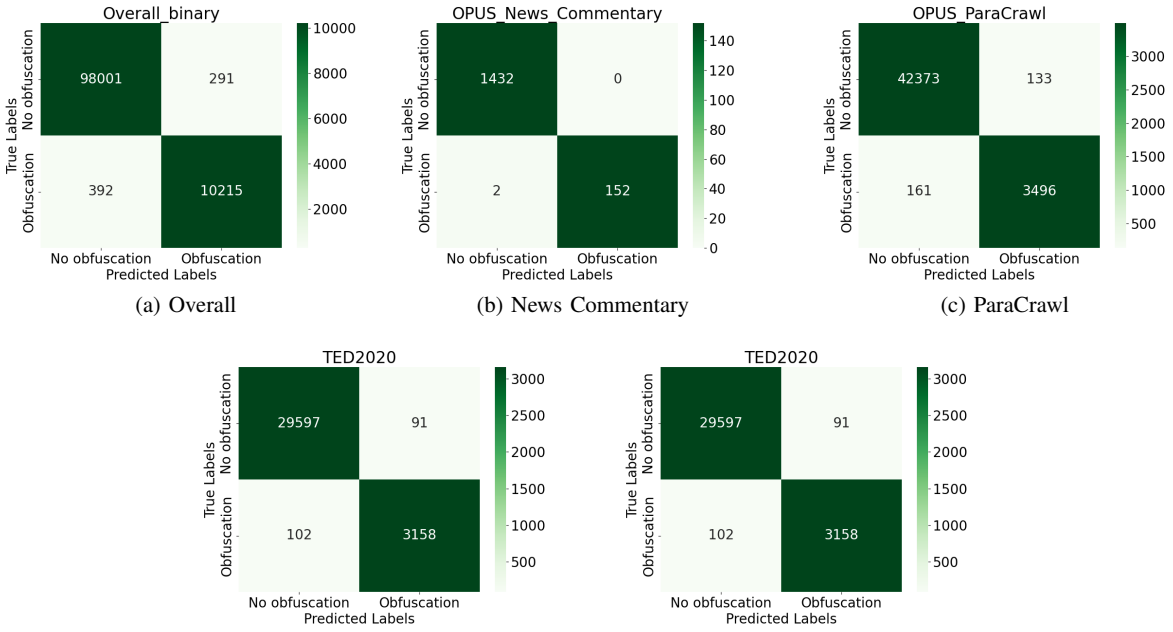


(d) TED2020



(e) WikiMatrix

Figure 1: Confusion matrices achieved by the different assessed Transformer models. The aggregated results for the multilingual test dataset is shown in (a), while the results for the separated datasets (i.e. News Commentary, ParaCrawl, TED2020, and WikiMatrix) are shown in (b)-(e), respectively. Rows indicate the actual camouflage types and columns show the predictions.



(a) Overall (b) News Commentary (c) ParaCrawl

(d) TED2020 (e) WikiMatrix

Figure 2: Binary confusion matrices for word camouflage detection by RoBERTa. (a) depicts the aggregated results for the multilingual test dataset, while the results for the separated datasets (i.e. News Commentary, ParaCrawl, TED2020, and WikiMatrix) are shown in (b)-(e). Rows indicate whether there is actually obfuscation or not and the columns show whether the model has predicted such obfuscation.

been effective. As a general rule, the models demonstrate very similar F1 scores, but **RoBERTa** stands out by outperforming the others in almost all metrics.

Additionally, we have compared the Transformer model that achieved the best performance (i.e., RoBERTa) with LeetSpeaker [5]. This comparison is shown in Table II, and they show that RoBERTa surpasses the Leetspeaker model, showing high weighted and micro F1 scores. RoBERTa has proved superior on these metrics, highlighting its capability in handling imbalanced class distributions effectively. However, the LeetSpeaker model consistently achieves higher macro F1 scores across datasets, indicative of strong performance on minority classes.

The confusion matrices (shown in Figure 1) show the results of evaluating the model on the subsets of the dataset. For the analysis, we will focus on the overall graphic which shows that the model correctly classifies 98,140 instances of class O (No obfuscation). The size of this number should not be taken into account when evaluating the model. Instead, the true performance should be evaluated by considering the ratio of correctly classified instances to the total instances per class. The other four classes —INV_CAMO, LEETSPEAK, PUNCT_CAMO, and MIX— show lower correct classifications due to the amount of data in each class. However, most of the words in these classes are correctly predicted. However, there are confusions between classes that should be noted. In particular, MIX is frequently misclassified as LEETSPEAK (793 instances), and PUNCT_CAMO is often confused with MIX (436 instances). This is probably due to the nature of the MIX class, as it is a combination of the LEETSPEAK and PUNCT_CAMO classes, which challenges the model's ability to establish clear boundaries between them.

Table II: F1 scores (macro, micro, weighted) comparing RoBERTa with the LeetSpeaker model from Huertas et al. [5]. The results of the Leetspeaker model have been extracted from [5]. The highest value for each dataset is shown in bold.

Dataset	F1	LeetSpeaker	RoBERTa
Overall	macro	0.8958	0,8783
	micro	0.9817	0.9835
	weighted	0.8126	**0.9835**
OPUS News Commentary	macro	0.9167	0.8977
	micro	0.9868	0.9868
	weighted	0.8515	**0.9862**
OPUS ParaCrawl	macro	0.8690	0.8485
	micro	0.9809	0.9842
	weighted	0.7750	**0.984**
TED 2020	macro	0.9224	0.8875
	micro	0.9854	0.9845
	weighted	0.8500	**0.9841**
WikiMatrix	macro	0.9032	0.8794
	micro	0.9785	0.9804
	weighted	0.8245	**0.9798**

On the other hand, Figure 2 reflects a significant improvement over the results of the previous one. This table shows how the model is able to correctly predict whether a word is obfuscated or not, which would meet the main objective.

V. Conclusions

In this work, we evaluated five different Transformer models: BERT, DistilBERt, MPNet, RoBERTa, XLNet to

detect obfuscation in text. We followed a Named Entity Recognition(NER) classification approach to detect the different obfuscation techniques and trained these models on the XX_NER_WordCamouflage training dataset.

This study aims to compare these models with the one proposed by Huertas et al. [5], with the goal of improving the performance of obfuscation detectors in real-world scenarios—such as content moderation. This will help to prevent crimes such as fraud, drug trafficking, or child sexual abuse.

Among these models, RoBERTa achieved the best performance. Based on this result, we selected RoBERTa and compared its results against those obtained by Huertas et al. [5], who also used Transformer models to detect obfuscated text.

The results show that RoBERTa achieved the best perfomance among the other evaluated Transformer models. Moreover, RoBERTa also outperformed the LeetSpeak model of Huertas et al. [5] in terms of weighted F1 scores, meaning that RoBERTa stands out for its strength in correctly predicting classes with the highest number of instances. In addition, we conclude that the most difficult class to detect was MIX due to the large number of confused samples.

Despite advancements in text obfuscation and detection, several challenges remain. For future work, we would like to evaluate Large Language Models (LLM) to detect obfuscated text and correct it. In this way, we will be able to positively influence the performance of automated content moderation systems.

The current dataset includes only four obfuscation techniques, which limits the models' effectiveness in real-world applications. To overcome this, future research will expand the dataset to cover a wider variety of obfuscation methods, enhancing the models' accuracy and adaptability.

Acknowledgements

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation) thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

References

[1] I. Ben Cheikh Larbi, A. Burchardt, and R. Roller, "Clinical Text Anonymization, its Influence on Downstream NLP Tasks and the Risk of Re-Identification," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, E. Bassignana, M. Lindemann, and A. Petit, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 105–111. [Online]. Available: https://aclanthology.org/2023.eacl-srw.11/

[2] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, "The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization," *Computational Linguistics*, vol. 48, no. 4, pp. 1053–1101, 12 2022. [Online]. Available: https://direct.mit.edu/coli/article-pdf/48/4/1053/2062009/coli_a_00458.pdf

[3] J. Pastor-Galindo, H.-Â. Sandlin, F. G. Mármol, G. Bovet, and G. M. Pérez, "A big data architecture for early identification and categorization of dark web sites," *Future Generation Computer Systems*, vol. 157, pp. 67–81, 2024.

[4] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: analysis of spammer strategies and the dataset shift problem," *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1145–1173, 2023.

[5] Álvaro Huertas-García, A. Martín, J. Huertas-Tato, and D. Camacho, "Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage," *Applied Soft Computing*, vol. 145, p. 110552, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623005707

[6] D. Nozza, D. Hovy *et al.*, "The state of profanity obfuscation in natural language processing scientific publications," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023.

[7] A. R. Perez, P. Rivas, J. Turek, K. Sooksatra, E. Quevedo, G. Bichler, T. Cerny, L. Giddens, and S. Petter, "Decoding the obfuscated: Advanced ner techniques on commercial sex advertisement data," in *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2023, pp. 144–151.

[8] K. Blashki and S. Nichol, "Game geek's goss: Linguistic creativity in young males within an online university forum (94/\/\3 933k'5 9055oneone)," *Australian Journal of Emerging Technologies and Society*, vol. 3, pp. 77–86, 2005.

[9] W. Peng, L. Huang, J. Jia, and E. Ingram, "Enhancing the naive bayes spam filter through intelligent text modification detection," in *2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE)*. IEEE, 2018, pp. 849–854.

[10] S. Renwick and D. Barbosa, "Detection and identification of obfuscated obscene language with character level transformers." in *Canadian AI*, 2021.

[11] Y. Sanjalawe, S. Fraihat *et al.*, "Detection of obfuscated tor traffic based on bidirectional generative adversarial networks and vision transform," *Computers & Security*, vol. 135, p. 103512, 2023.

[12] W. Zhu and S. Bhat, "Euphemistic phrase detection by masked language model," *arXiv preprint arXiv:2109.04666*, 2021.

[13] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, and A. Panchenko, "Paradetox: Detoxification with parallel data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6804–6818.

[14] E. Xing, S. Venkatraman, T. Le, and D. Lee, "Alison: Fast and effective stylometric authorship obfuscation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 315–19 322.

[15] D. Noever, J. Kalin, M. Ciolino, D. Hambrick, and G. Dozier, "Local translation services for neglected languages," 2021. [Online]. Available: https://arxiv.org/abs/2101.01628

[16] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, and A. Panchenko, "Text detoxification using large pre-trained neural models," *CoRR*, vol. abs/2109.08914, 2021. [Online]. Available: https://arxiv.org/abs/2109.08914

[17] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, and A. Panchenko, "ParaDetox: Detoxification with parallel data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6804–6818. [Online]. Available: https://aclanthology.org/2022.acl-long.469

[18] W. Zhu and S. Bhat, "Euphemistic phrase detection by masked language model," *CoRR*, vol. abs/2109.04666, 2021. [Online]. Available: https://arxiv.org/abs/2109.04666

[19] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han, "Automated phrase mining from massive text corpora," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1825–1837, 2018.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020. [Online]. Available: https://arxiv.org/abs/1910.01108

[23] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," 2020. [Online]. Available: https://arxiv.org/abs/2004.09297

[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2020. [Online]. Available: https://arxiv.org/abs/1906.08237

A review of DeepHSAR: semi-supervised fine-grained learning for multi-label human sexual activity recognition

Alicia Martínez-Mendoza* amartm@unileon.es
Enrique Alegre* ealeg@unileon.es
Eduardo Fidalgo* efidf@unileon.es
Christopher Gaul* cgau@unileon.es
Irene Aguado-Caballero* iaguc@unileon.es

*Department of Electrical, Systems and Automation Engineering, Universidad de León Campus Vegazana, 24007 León, Spain

***Abstract*—The identification of sexual activities in images is crucial for content severity classification and improving pornography detection systems. This study introduces DeepHSAR, a semi-supervised deep learning framework designed for fine-grained, multi-label Human Sexual Activity Recognition (HSAR). The study also presents SexualActs-150k, a dataset containing 150k images labeled with 19 types of sexual activities. DeepHSAR employs two classification streams: (i) a global stream for coarse-grained classification and (ii) a local stream leveraging semi-supervised clustering and CNN training to recognize fine-grained visual details. The framework achieves an F1-score of 79.29%, outperforming existing multi-label classification approaches. Additionally, DeepHSAR establishes new state-of-the-art results on Pornography-2k, MS-COCO, and NUS-WIDE datasets. This contribution is significant in the cybersecurity field for law enforcement agencies (LEAs) and pornography filtering systems, enabling automated categorization of explicit content and enhanced recognition of high-risk material in digital forensics.**

***Index Terms*—cybersecurity, multi-label classification, sexual activity detection, fine-grained classification, semi-supervised classification, pornography detection**

Type of contribution: *Already published research [1]*

I. Introduction

With the widespread availability of the Internet, the production and sharing of Child Sexual Exploitation Material (CSEM) has drastically increased, presenting a significant challenge for Law Enforcement Agencies (LEAs). Therefore, this research [1] addresses the need for automated systems that detect and categorize sexual activities in images, potentially improving CSEM detection and content moderation.

We propose the task Human Sexual Activity Recognition (HSAR). Its objective is to determine the obscene content severity in the pornographic or CSEM material. We also introduce SexualActs-150k, the first publicly available multi-label dataset for HSAR, containing 150k images with 19 categories representing sexual activities. In addition, we present Deep Human Sexual Activity Recognition (DeepHSAR), a novel deep learning-based framework for detecting sexual activities in images, together with a semi-supervised learning method that learns discriminative patches in the images through iterative clustering and CNN learning.

Our proposed framework aims to identify distinct semantic regions associated with sexual activity as individual action components, extract critical activity-related cues from these regions, and integrate this information to facilitate multi-label classification, enabling the detection of all sexual activities present in an image. Moreover, to summarize the sexual activities, we exploit "sexual activity-specific semantic regions" through "visual words" (e.g., penis, vulva) and "visual phrases" (e.g., kissing, fellatio).

II. SexualActs-150k Dataset

SexualActs-150k contains 150,720 images, each labeled with all the sexual activities present in the image. A total of 19 categories has been chosen based on content severity and on the categories in the Combating Paedophile Information Networks in Europe (COPINE) and Sentency Advisory Panel (SAP) scales, including, for example, *Indicative*, *Kissing*, *Vulva Hand Sex*, *HandJob*, *Toy Sex*, *Vaginal Penetration*, *BDSM*, and *Bestiality*. The images have been crawled from Google and DuckDuckGo search engines from pornographic websites.

III. Proposed Approach

For the image representation backbone, we adopted the ResNet-101 model and added an attention module [2]. We pretrained this network, referred to as ResAttNet-101_pre, using cross-entropy loss on the pornography vs. safe classification task using the large-scale Pornographic-2M dataset [3].

Our proposed framework consists of two streams:

- The **Global Feature Stream**, which uses a modified ResAttNet-101_pre network. We incorporated an auxiliary output layer and replaced the cross-entropy loss with a multi-label prediction head, creating the ML-ResAttAuxNet-101 network.
- The **Local Feature Stream**, whose objective is to perform multi-label and fine-grained image classification. We created a pipeline that uses an iterative clustering and a CNN fine-tuning process using labeled and unlabeled sets.

We fused both streams within our proposed framework to leverage global semantic abstraction and features from local discriminative regions. To accomplish this fusion, we utilized two different strategies: at feature level and at score level.

A. Evaluation protocol

To evaluate our proposed framework, we selected metrics widely adopted in multi-label image classification: average

TABLE I
PERFORMANCE RESULTS FOR DEEPHSAR AND STATE-OF-THE-ART MODELS ON THE SEXUALACTS-150K DATASET. DEEPHSAR VERSIONS: G = GLOBAL FEATURE STREAM, L = LOCAL FEATURES STREAM, S = SCORE FUSION OF 2 STREAMS, AND F = FEATURE FUSION OF 2 STREAMS.

Model	P_C	R_C	$F1_C$	P_O	R_O	$F1_O$
KSSNet [4]	64.45	57.67	60.87	74.95	66.38	70.40
C-Tran [5]	68.23	62.41	65.19	77.125	68.09	72.32
ResNet-101-BCE	64.16	57.25	60.51	74.31	66.13	69.98
DeepHSAR-G	67.34	59.72	63.30	75.53	66.89	70.95
DeepHSAR-L	71.98	64.35	67.95	80.52	71.30	75.63
DeepHSAR-S	**78.94**	**69.95**	**74.17**	**84.42**	**74.75**	**79.29**
DeepHSAR-F	78.35	69.46	73.63	83.96	74.35	78.86

precision, recall, and F1-score of the predicted image labels, and we calculated them both per class ($P_C, R_C, F1_C$) and overall ($P_O, R_O, F1_O$).

B. Baseline and State-of-the-art Methods

We performed a comparative analysis with seven state-of-the-art methods including Multi-Label Image Recognition with Graph Convolutional Networks (MLGCN), Multi-Label Classification with Label Graph Superimposing (KSSNet) [4] and Classification Transformer (C-Tran) [5] among others [1]. We also computed results for ResNet101 with Binary Cross Entropy (BCE). The comparison between these methods is shown in Table I.

C. Experiments on Benchmark Datasets

We evaluated our proposed framework on the SexualActs-150k dataset (as mentioned in Section III-B), and in two standard multi-label classification benchmark datasets: NUS-WIDE [6] and MS-COCO [7], obtaining comparative results with state-of-the-art models. In addition, we tried our model on the Pornography-2k dataset, which contains videos of pornographic and non-pornographic scenes for binary classification. DeepHSAR achieves an accuracy of 99.85%, outperforming the other five state-of-the-art approaches, as shown in Table II.

The experiments on the first three datasets indicated that the local stream provided better results than the global feature stream. Moreover, the performance improved with both fusion strategies of the streams. The performance obtained by the score-level fusion was higher than that obtained by the feature-level fusion.

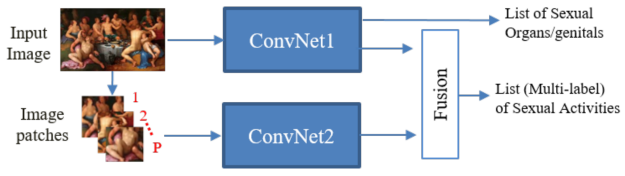


Fig. 1. Summary of the model's structure

TABLE II
RESULTS FOR VIDEO CLASSIFICATION ON NPDI PORNOGRAPHY-2K DATASET

Model	**Accuracy.(%)**
ResAttNet-101_pre	98.78
ODT [37]	98.8
DVRGNet [36]	99.04
Gautam et al. [13]	97.15
Samal et al. [38]	99.15
DeepHSAR (ours)	**99.85**

IV. CONCLUSION

We introduced a novel task named automated Human Sexual Activity Recognition and presented SexualActs-150k, a new multi-label sexual activity dataset designed to evaluate this task. We proposed DeepHSAR, a framework that addresses fine-grained multi-label image classification. It consists of two streams designed to extract global and local features.

We evaluated our framework and seven state-of-the-art methods using the SexualActs-150k, MS-COCO, NUS-WIDE, and Pornography-2k datasets. The results showed that the proposed DeepHSAR approach obtained remarkable performance compared with other state-of-the-art methods.

The capabilities of our proposal could be beneficial for a variety of applications in cybersecurity, including assisting LEAs in detecting CSEM. DeepHSAR's multi-label classification provides a more granular analysis than binary classification, offering information about the severity of the content and potentially aiding in prioritization tasks.

ACKNOWLEDGMENTS

This work has been supported thanks to the Plan of Recovery, Transformation and Resilience, financed by the European Union (Next Generation) thanks to the LUCIA Project (Spanish acronym meaning Fight against Cybercrime using Artificial Intelligence), granted to the University of León by INCIBE. The research presented in this work was funded with support from the European Union's Horizon 2020 Research and Innovation Framework Programme, H2020 SUFCT-2019 under the GRACE project with Grant Agreement 883341. This publication reflects the views only of the authors, and the European Union's Horizon 2020 Research and Innovation Framework Programme, H2020 SUFCT-2019 cannot be held responsible for any use that may be made of the information contained therein.

REFERENCES

[1] Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E. & Martínez-Mendoza, A. (2024). DeepHSAR: semi-supervised fine-grained learning for multi-label human sexual activity recognition. Information Processing & Management, 61, 103800.
[2] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., et al. (2017). Residual attention network for image classification. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (pp. 6450–6458).
[3] Gangwar, A., González-Castro, V., Alegre, E., & Fidalgo, E. (2021). AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images. Neurocomputing, 445, 81-104.
[4] Wang, Y., He, D., Li, F., Long, X., Zhou, Z., Ma, J., & Wen, S. (2019). Multi-label classification with label graph superimposing. In AAAI Conference on Artificial Intelligence.
[5] Lanchantin, J., Wang, T., Ordonez, V., & Qi, Y. (2021). General Multi-label Image Classification with Transformers. In CVPR.
[6] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: a real-world web image database from National University of Singapore. In CIVR '09.
[7] Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In ECCV.

Detección de vulnerabilidades en TensorFlow comparando redes BERT y LSTM

Sergio Muñoz Martín, Luis Alberto Martínez Hernández, Sandra Pérez Arteaga,
Ana Lucila Sandoval Orozco, Luis Javier García Villalba*

Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA)
Facultad de Informática, Despacho 431, Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid
Email: {sergmu03, luisal11, sandrp08, asandov, javierv}@ucm.es

***Resumen*—En este trabajo se ha desarrollado un modelo de detección de vulnerabilidades en código Python en los commits de TensorFlow, con el objetivo de mejorar la seguridad de los datos y facilitar el reconocimiento de errores en el desarrollo de aplicaciones basadas en inteligencia artificial. Para ello, se han implementado y comparado dos enfoques de Aprendizaje Profundo: uno implementando una red neuronal LSTM, especializada en la captura de dependencias secuenciales en el código, y otro que emplea un modelo BERT adaptado al análisis semántico del texto. Esta investigación no solo busca fortalecer la seguridad en el uso de este framework, sino también sentar las bases para su futura aplicación en otros entornos de desarrollo de Aprendizaje Profundo como Pytorch. De esta manera, se pretende proporcionar una herramienta que ayude a los desarrolladores en la detección temprana de vulnerabilidades, fomente la escritura de código seguro y robusto, y contribuya a la integración de técnicas de procesamiento del lenguaje natural en el análisis automatizado de software. Las pruebas realizadas son positivas, los dos modelos alcanzan en las pruebas test una puntuación F1 superior al 80 % y BERT ha demostrado ser un modelo superior para esta tarea.**

***Index Terms*—Análisis estático, Aprendizaje Profundo, Ciberseguridad, BERT, LSTM, TensorFlow, Vulnerabilidades**

Tipo de contribución: *Investigación original*

I. Introducción

La tecnología se ha vuelto un pilar esencial para la vida cotidiana de las personas y para el funcionamiento de las empresas, y con ello la seguridad de los sistemas informáticos ha adquirido un papel crucial para proteger la información, especialmente en la seguridad del software en el que las vulnerabilidades suelen originarse en el código, representando una amenaza potencial para la integridad de los sistemas. Los algoritmos de Inteligencia Artificial (IA) se han vuelto muy populares y se están comenzando a implementar en la mayoría de ámbitos laborales, trabajando con datos sensibles como los datos financieros de un cliente por lo que esta investigación pretende mejorar la seguridad que hay actualmente al trabajar con algoritmos de IA.

Para proteger esta información es importante garantizar la protección de la Tríada de la Seguridad Informática (Confidencialidad, Integridad y Disponibilidad). Una vulnerabilidad en cualquiera de estos pilares puede tener consecuencias económicas, legales y sociales y para evitar esto además de la implementación de robustos protocolos de seguridad, es necesario que se añadan mecanismos de seguridad en el código como el análisis de código estático y dinámico para poder proteger la información de todas las amenazas.

La OWASP [1] realizó en 2023 un estudió en el que se recogen en orden las principales vulnerabilidades y riesgos de seguridad que aparecen en Modelos de Lenguaje Extensos (LLMs) y tecnologías derivadas según su criterio. El top 10 está formado por Ataque de manipulación de entrada, ataque de envenenamiento de datos, ataques de inversión del modelo, ataque de inferencia de membresía, robo de modelo, ataques a la cadena de suministro, ataque de aprendizaje por transferencia,modelo sesgado, ataque de integridad de salida, envenenamiento del modelo.

De todas estas amenazas que recoge la OWASP esta investigación se centra en los **Ataques a la cadena de suministro.** Estos ataques apuntan a los paquetes y frameworks que se utilizan para desarrollar los modelos, en los que ya viene incluido el código necesario para hacer el preprocesamiento de los datos y el entrenamiento de las redes neuronales.

Entre los que se utilizan en la cadena de suministro en el desarrollo de aplicaciones de IA en este trabajo se ha elegido Tensorflow [2] uno de los Frameworks de IA más utilizados en Python y que incluye en su paquete la gran mayoría de redes con las que se trabajan hoy en día.Esta investigación no solo busca mejorar la seguridad de Tensorflow sino que también pretende pretende ser una solución para el resto de frameworks de Aprendizaje profundo en un futuro y ayudar a los desarrolladores a encontrar vulnerabilidades ya existentes para facilitar la escritura de código seguro.

El resto del trabajo se organiza de la siguiente manera: en la Sección II se presenta el contexto del trabajo explicando conceptos como las redes que se van a utilizar y la seguridad en el código. En la Sección III se presentan las propuestas encontradas similares a este trabajo de investigación. La metodología seguida para el desarrollo del trabajo se muestra en la Sección IV. En la Sección V se muestran los experimentos y resultados y, finalmente, en la Sección VI se muestran las conclusiones del trabajo.

II. Contexto del trabajo

II-A. Redes Neuronales Recurrentes LSTM

Las Redes Neuronales Recurrentes (RNN, del inglés Recurrent Neural Networks) [3] son un grupo de redes neuronales que destacan por realizar un procesamiento de los datos de

manera secuencial. Debido a este entrenamiento secuencial en el que las salidas se calculan con los estados de los datos anteriores, se suele decir que las RNN tienen una memoria interna. A pesar de tener memoria interna, se dice que las RNN no tienen memoria a largo plazo. Como en este trabajo se emplean grandes cantidades de datos para el entrenamiento se va a utilizar específicamente una red LSTM [4]. Este tipo de red es conocida por ser una RNN con memoria de largo plazo, en la que se soluciona el problema de la memoria interna conocido como desvanecimiento del gradiente.

II-B. BERT

Cuyas siglas vienen del término en ingles, Bidirectional Encoder Representations from Transformers [5], es un modelo de aprendizaje profundo basado en la arquitectura Transformers. Este modelo ha sido diseñado y entrenado en la representación del lenguaje por lo que es un modelo ideal para tareas que necesitan una compresión del contexto del texto. Además BERT es un modelo bidireccional es decir analiza el texto en los dos sentidos y lanza sus predicciones en función de lo que haya tanto delante como atrás lo que le permite alcanzar un mayor rendimiento en el análisis de texto o en la clasificación de código en el que va a ser utilizado.

III. Estado del arte

Al entrenar un modelo para detectar vulnerabilidades se puede presentar el código de maneras muy diferentes. Empezando por entregar el código completo para que sea la red la que analice si existe una vulnerabilidad en ese fichero. En la propuesta [6] donde se realizaron experimentos en datasets del lenguaje C y C++ con redes GNN, modelos basados en RoBERTa o modelos basados en T5, sin llegar a superar una puntuación de F1 del 50 %. Por esto mismo, la mayoría de las propuestas del estado del arte realizan un preprocesamiento de los datos con el objetivo de ayudar al modelo a encontrar las relaciones que existen en el código para que sea vulnerable o no. Una metodología de preprocesamiento con los datos es utilizar las redes GNN capaces de procesar directamente datos estructurados en forma de grafo y aprender de las relaciones entre los nodos. Al utilizar estas redes, modelos como FUNDED [7], Mvulpreter [8] y Vulccn [9] buscan generar un grafo en el que las variables del código, objetos y funciones aparezcan representados mediante vértices y se utilicen los enlaces o ramas para representar sus relaciones y dependencias. El uso de esta metodología permite recoger de mejor manera las relaciones que aparecen en el código y la tarea de encontrar qué puntos son positivos o negativos se vuelve más fácil para la red [7].

El resto de propuestas realizan la división del código en partes o secuencias según diferentes criterios. El objetivo que se persigue con esta idea es el de entregarle a la red trozos importantes del código y que sea la red la que realice todo el análisis del lenguaje natural. Su función principal es recorrer todo el código generando bloques pequeños de código mediante una ventana de m caracteres que contiene el contexto al analizar.

Uno de los puntos negativos que tienen estas propuestas es que al analizar el código en secuencias de datos contiguas de un tamaño limitado, se puede dar el caso en el que la secuencia no recoja todo el contexto que afecta a esa zona del código, perdiendo en el análisis dependencias que pueden afectar a la vulnerabilidad. Sin embargo, en [10] y [11] argumentan que al estar escrito por humanos, el código tiene muchas similitudes con el lenguaje natural, ya que los humanos, al repetir tareas como la escritura de código, solemos desarrollar ciertos patrones y repetir estructuras que vamos desarrollando durante el aprendizaje de la tarea.

Tabla I
Resumen Estado del arte

Modelo	Método	Lenguage	Red	F1
DiverseVul	Fichero	C y C++	GNN	29.76 %
DiverseVul	Fichero	C y C++	RoBERTa	37.85 %
DiverseVul	Fichero	C y C++	GPT-2	35.23 %
DiverseVul	Fichero	C y C++	T5	47.15
FUNDED	Grafos	C, Java, PHP	GNN	94 %
mVulPreter	Slices	C y C++	GNN	78.8 %
VulDeePecker	C.Gadgets	C y C++	BLSTM	85.5 %
μvuldeepecker	C.Gadgets	C y C++	BLSTM	94.22 %
Vudenc	Secuencial	Python	LSTM	80 %
BBVD	Secuencial	C y C++	RoBERTa	89.37

Entre las propuestas que analizan el código dividiéndolo en secuencias se encuentra: Vudenc [11] que utiliza una red LSTM para analizar funciones de código C dividiendo el código en bloques y se obtienen F1 del 80 % con el algoritmo RMSprop. μvuldeepecker [12] y VulDeePecker [13] proponen variaciones a los modelos anteriores para superar los puntos negativos. Sus propuestas tratan de agrupar las zonas que se consideran que afectan a la vulnerabilidad o puntos de atención (mediante esta metodología, un bloque de datos de entrenamiento puede contener datos de líneas de texto diferentes). Sin embargo, con los puntos de atención se puede perder información relevante para una red cuyo objetivo es analizar el lenguaje natural encontrando patrones en el código. Las mejores métricas de F1 obtenidas en estas propuestas fueron 94.22 % y 85.5 %. Finalmente, también se ha encontrado BBVD[14] en la que se prueba un modelo de secuencia para lenguajes C y C++ con redes basadas en BERT.

En la tabla I aparece un resumen de las métricas sacadas en cada artículo centrándose en las métricas F1 que presentan en sus artículos.

IV. Metodología

Para este trabajo se ha elegido probar dos modelos. Primero BERT debido a su desempeño probado en este tipo de tareas en investigaciones recientes. Y Finalmente una red LSTM que como se ha visto en la propuesta [11] ya se ha probado que pueden ofrecer buenos resultados en estas tareas y además requieren un menor carga computacional y son mucho más ligeras que las LLM que están siendo tan populares recientemente.

En la Figura 1 se muestra el diagrama general de la propuesta, el cual consta de 4 pasos necesarios hasta generar el modelo.

IV-A. Recolección de los datos

Esta etapa es el primer cuadro de la Figura 1 en la que el objetivo es encontrar y almacenar un conjunto de datos con vulnerabilidades que permita al modelo extraer la información necesaria para encontrar vulnerabilidades después del entrenamiento.

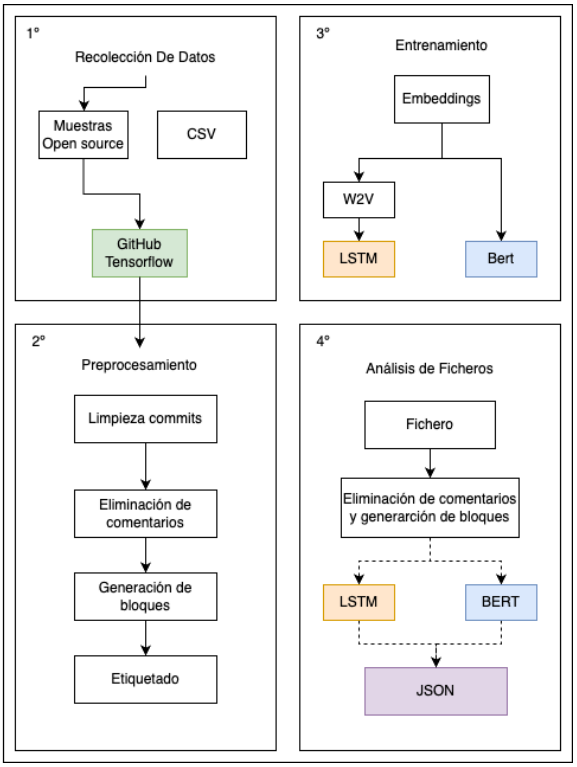


Figura 1. Diagrama general de la Propuesta

TensorFlow es un framework que está desarrollado principalmente en C++, pero también tiene código en Python para que pueda ser utilizado en este lenguaje de programación. Al buscar conjuntos de datos o estudios sobre las vulnerabilidades recogidas en TensorFlow se observa como la mayoría de los artículos [15], [16] se centran en el código C++, y recogen vulnerabilidades clásicas como desbordamiento de enteros o la escritura fuera de límites.

Por lo tanto, para obtener un conjunto de datos con vulnerabilidades en Python se han recogido vulnerabilidades de 141 commits registrados en snyk [17], una plataforma permite a los desarrolladores de código y equipos de DevOps [18] identificar, solucionar y monitorizar vulnerabilidades en sus aplicaciones y dependencias a lo largo del ciclo de vida del desarrollo de software, por lo que resulto ideal para realizar una búsqueda inicial de las vulnerabilidades que han ido apareciendo en TensorFlow. Sin embargo como la gran mayoría de commits registrados afectaban principalmente al código C de TensorFlow fue necesario buscar mas muestras de datos. Para aumentar el tamaño del conjunto de datos se extrajeron del repositorio de Github de TensorFlow [2] hasta 1500 commits, a través de una búsqueda en la que se extrajeron los commits que estaban relacionados con vulnerabilidades o la resolución de bugs identificando commits en los que el mensaje tiene palabras como: vulnerability, security, danger, bug, exploit, cve, etc.

Esta metodología de recolección de los datos es diferente a la mayoría del Estado del Arte donde se basan en vulnerabilidades que han sido reportadas y recopiladas en bases de datos donde se adjunta toda la información sobre la vulnerabilidad. Sin embargo, esta propuesta se basa en las vulnerabilidades que han sido reportadas por los desarrolladores a lo largo de las diferentes versiones de TensorFlow. Al seguir este método, se tiene menos información sobre las vulnerabilidades que han sido recopiladas, pero como contrapunto se entrena el modelo en vulnerabilidades que no serían tratadas en otros conjuntos de datos más convencionales.

Algunos de los ejemplos de los mensajes de los commits recopilados:

- Fix security vulnerability with FractionalMax(AVG)Pool with illegal pooling_ratio
- Security vulnerability fix for Out of Bound error in ThreadUnsafeUnigramCandidateSampler.

IV-B. Preprocesamiento de los datos

Esta etapa se realizan los siguientes pasos recogidos en el bloque 2 de la Figura 1 en los que se modifican los datos hasta conseguir el mejor formato posible para el entrenamiento de la red: limpieza de código, eliminación de comentarios, división del código en bloques, etiquetado de bloques.

IV-B1. Limpieza de los datos: En la limpieza de datos primero se revisan y eliminan aquellos commits cuyas modificaciones afecten solo a comentarios del código, aquellos commits que se encuentren repetidos y se eliminan todos los comentarios que no afectan a la ejecución de los ficheros de código, ya que no influyen en la lógica y comportamiento final del programa.

IV-B2. División del código en bloques y etiquetado: Para obtener un modelo capaz de identificar qué zona del código es vulnerable se va a analizar cada línea de código de manera secuencial. Como el objetivo del modelo no es solo identificar si el código es vulnerable, sino la zona en la que se encuentra la vulnerabilidad, es necesario entrenar el código por partes, como se vio en el Estado del Arte. En este caso se va a dividir de manera secuencial el código en bloques. Para determinar el tamaño que van a tener los bloques de código se pueden implementar distintas medidas de referencia como los caracteres, los tokens o las líneas de código. En este caso se han elegido las líneas de código, ya que coincide con las líneas eliminadas del commit de Github. Cada bloque contendrá una línea de código que se va a analizar, y que estará acompañada de n líneas de contexto por delante y detrás que permiten al modelo analizar las relaciones y patrones que aparecen en el código para determinar si es vulnerable o no. El funcionamiento de este procedimiento se puede observar en la Figura 2. En esta figura, aparece un fragmento de código vulnerable y se procede a obtener los bloques con 2 líneas de contexto por delante y por detrás. Así, en el bloque verde las dos primeras líneas del código son de contexto, la línea 3 es la línea principal que se va a analizar si es vulnerable o no y las líneas 4 y 5 forman el contexto posterior del bloque. Luego de manera secuencial se forman el resto de bloques en el bloque 2 la línea principal sería la 3, en el 3 la 4, hasta recorrer todo el fichero de código.

Además, a la vez que se está generando cada bloque, se utiliza la información de las líneas eliminadas del commit para comprobar si la línea que es el centro de atención es vulnerable o no. Si la línea coincide con una de las líneas eliminadas, se etiqueta el bloque como 1 o 0 en caso contrario.

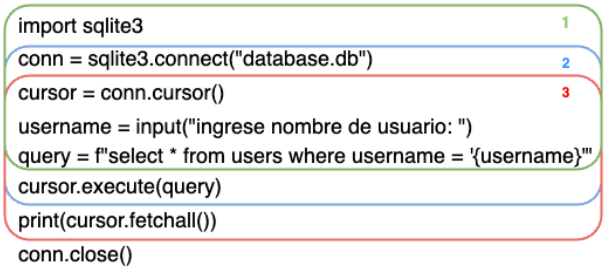


Figura 2. Diagrama de bloques de la figura

IV-C. Entrenamiento del modelo

Para que las redes neuronales sean capaces de interpretar el lenguaje del código de programación es necesario realizar una transformación del código en vectores numéricos conocidos como Embeddings. Estos embeddings se generan mediante el entrenamiento de una red neuronal, que estima para cada palabra la probabilidad de encontrar cada una del resto de palabras del vocabulario en su contexto. Con esto, lo que se consigue es que dos palabras semánticamente similares tengan embeddings similares. Como se observa en el bloque 3 de la figura 2, para generar los embeddings de la red LSTM se ha utilizado la biblioteca de la red Word2Vec. Por otro lado, BERT incorpora un mecanismo de embeddings basado en la arquitectura Transformer y se encarga de generar automáticamente representaciones contextuales potentes durante su procesamiento.En la etapa de entrenamiento se utiliza el código transformado en vectores numéricos y su etiquetado para entrenar los modelos LSTM y BERT. Además para procesar la salida de los modelos y que puedan ser utilizados para tareas de clasificación binaria es necesario utilizar agregar a la salida una capa adicional que procese la información. En la red LSTM se agregó una capa Densa con Sigmoid, mientras que en BERT fue necesario agregar una capa argmax.

IV-D. Análisis de ficheros

Tras finalizar el entrenamiento del modelo y determinar la configuración óptima, se utiliza el modelo entrenado para analizar ficheros de código e identificar qué líneas de código son vulnerables. Para ello, de manera muy similar a la que se hacía en el bloque 2 de la figura 1 se eliminan los comentarios del código y se recorre el código con un bucle para dividir el código en bloques. Sin embargo, en este caso, cada bloque se pasa a embedding y se utiliza el modelo BERT o LSTM para realizar predicciones e identificar si la línea central es vulnerable o no.

V. Experimentos y Resultados

Para probar los resultados del modelo y su rendimiento se utilizó un ordenador con CPU IntelCore i7-7920HQ a 3,10 GHz, memoria RAM de 16,0 GB y una GPU GForce RTX 3070 OC Edition con 8 GDDR6.

Para garantizar una evaluación rigurosa y equilibrada del modelo que permita evaluar la variabilidad del modelo, los experimentos se llevaron a cabo mediante una validación cruzada de 5 folds.

Tabla II
Resultados Experimentos

Model	Context	Accuracy	Precisión	Recall	F1
BERT	3 líneas	0.8751	0.8769	0.8751	0.8756
BERT	5 líneas	0.8969	0.8970	0.8969	0.8970
LSTM	3 líneas	0.8308	0.7562	0.8526	0.8012
LSTM	5 líneas	0.8430	0.7778	0.8508	0.8125

En estos experimentos en la red LSTM la tasa de aprendizaje elegida fue de 1e-5, para evitar el sobreajuste se estableció un dropout de 0.2 y se implementó un earlyStopping tomando como métrica de referencia la puntuación de F1, el batch size se estableció en 64 y se estableció un máximo de 30 épocas para cada folds. Para optimizar el entrenamiento ajustando la tasa de aprendizaje, se utilizó el algoritmo ADAM.

En el caso del modelo BERT se utilizó la versión BERT-base-uncased que consta de aproximadamente 110 millones de parámetros,se fijó una tasa de aprendizaje de 2e-5, para evitar el sobreajuste un weight decay de 0.01, el batch size se estableció en 16 y de igual manera que en el modelo LSTM se agregó un earlyStopping con un número máximo de 30 épocas.

Los resultados de los modelos entrenados se pueden observar en en la Tabla II en la que aparecen para cada modelo las métricas finales de test del cross-validation. Las métricas Accuracy, Precisión, Recall y F1 son las utilizadas por la mayoría de las propuestas del arte y se encuentran explicadas en [11]. Además en la Figura 3 se representa la evolución de la métrica accuracy que han tenido los modelos en una de las particiones del cross-validation.

En las gráfica se observa que el entrenamiento de modelo BERT en las dos versiones ha requerido un menor número de épocas, debido a la implementación del earlyStopping que previene el sobreentrenamiento. Desde el comienzo del entrenamiento el modelo BERT logra un mejor rendimiento en los datos de validación y se observa como al aumentar el contexto de los bloques y pasar de 3 lineas a 5 lineas el rendimiento mejora ligeramente tanto en el modelo BERT como en el LSTM.

El modelo BERT con tres líneas ha obtenido unas métricas positivas cercanas a 0,875 en todas las métricas y sin observar una reducción en la Precisión y en la F1 como sí se observaba en algunas propuestas del estado del arte, lo que indica que el modelo tiene un rendimiento sólido en los experimentos realizados. Al agregar más contexto en el modelo de 5 líneas, el modelo aumenta sus métricas a 0.897, por lo que se observa que aumentar el contexto ayuda al modelo a encontrar los patrones que hacen que un bloque de código sea vulnerable o no.Por otra parte, en el modelo LSTM, al ser un modelo inferior, era previsible que las métricas disminuyeran y así ha sido, llegándose a obtener una puntuación F1 de 0.80 en el modelo de tres líneas y 0.81 en el modelo de 5 líneas. Comparando los modelos LSTM y BERT se observa que el aumento de las métricas en la red LSTM al aumentar el contexto es inferior al modelo BERT, y las métricas de precisión disminuyen respecto al resto de métricas considerablemente en el modelo LSTM.

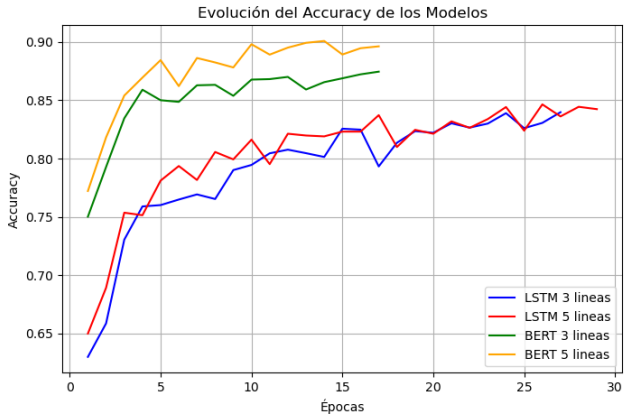


Figura 3. Evolución Accuracy

VI. Conclusiones

Este trabajo ha comenzado el desarrollo de una herramienta que ayude a los programadores a identificar vulnerabilidades en Python al trabajar en algoritmos de Inteligencia Artificial. Las pruebas realizadas hasta el momento en el framework TensorFlow son positivas y prometedoras, demostrando que estas herramientas son capaces de implementarse para la detección de fallos de seguridad en este ámbito y muchos otros como se ha visto en el estado del arte. Los experimentos realizados demuestran que BERT es más efectivo que LSTM para esta tarea, especialmente cuando se incrementa la cantidad de contexto. La capacidad de BERT para generar embeddings contextuales robustos se refleja en sus mejores resultados, mientras que el LSTM, aunque mejora con más contexto, no es suficiente en comparación con el modelo BERT por lo que aunque requiere una capacidad de computación menor, es necesario aumentar su rendimiento para que sea viable en un ambiente real. Uno de los problemas más relevantes que han aparecido durante la investigación es la recopilación de muestras de código Python vulnerable específicamente relacionadas con los algoritmos de inteligencia artificial. Esta escasez de muestras vulnerables puede afectar a la generalización del modelo por lo que para tratar este inconveniente en los próximos pasos de la investigación se propone complementar el conjunto de datos con vulnerabilidades de otros frameworks ya mencionados como pueden ser Torch, MxNet, o Hugging Face Transformers. Además, se puede complementar el conjunto de datos con vulnerabilidades más generales del lenguaje de Python para que reconozca también cosas como las inyecciones de código y de SQL, problemas de exposición de información sensible o dependencias vulnerables, entre otras.

Agradecimientos

Este trabajo ha sido apoyado por la Comisión Europea dentro del programa de investigación e innovación Horizonte Europa, como parte del proyecto LAZARUS (Grant Agreement nº 101070303). Este trabajo ha sido financiado también con fondos del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation), a través de la Cátedra INCIBE-UCM ≪Ciberseguridad para la Innovación y Protección Digital≫ (https://cybersec.ucm.es). No obstante, las opiniones expresadas son responsabilidad exclusiva de los autores y no reflejan necesariamente las de INCIBE, la Unión Europea o la Comisión Europea-UE. Ni INCIBE, ni la Unión Europea, ni la Comisión Europea pueden ser consideradas responsables de las mismas.

Referencias

[1] OWASP Foundation, "OWASP Machine Learning Security Top 10," https://owasp.org/www-project-machine-learning-security-top-10/, 2025, Último acceso: 14 de enero de 2025.

[2] T. Developers. (2024) Tensorflow: An open source machine learning framework for everyone. Último acceso: 4 de enero de 2025. [Online]. Available: https://github.com/tensorflow/tensorflow

[3] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 07 2019. [Online]. Available: https://doi.org/10.1162/neco_a_01199

[4] "Understanding LSTM Networks," https://colah.github.io/posts/2015-08-Understanding-LSTMs/, January 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Y. Chen, Z. Ding, L. Alowain, X. Chen, and D. Wagner, "Diversevul: A new vulnerable source code dataset for deep learning based vulnerability detection," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 2023, pp. 654–668.

[7] H. Wang, G. Ye, Z. Tang, S. H. Tan, S. Huang, D. Fang, Y. Feng, L. Bian, and Z. Wang, "Combining graph-based learning with automated data collection for code vulnerability detection," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1943–1958, 2021.

[8] D. Zou, Y. Hu, W. Li, Y. Wu, H. Zhao, and H. Jin, "mvulpreter: A multi-granularity vulnerability detection system with interpretations," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–12, 2022.

[9] Y. Wu, D. Zou, S. Dou, W. Yang, D. Xu, and H. Jin, "Vulcnn: An image-inspired scalable vulnerability detection system," in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022, pp. 2365–2376.

[10] R. L. Russell, L. Y. Kim, L. H. Hamilton, T. Lazovich, J. A. Harer, O. Ozdemir, P. M. Ellingwood, and M. W. McConley, "Automated vulnerability detection in source code using deep representation learning," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 757–762, 2018.

[11] L. Wartschinski, Y. Noller, T. Vogel, T. Kehrer, and L. Grunske, "Vudenc: Vulnerability detection with deep learning on a natural codebase for python," *Information and Software Technology*, vol. 144, p. 106809, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584921002421

[12] D. Zou, S. Wang, S. Xu, Z. Li, and H. Jin, "μvuldeepecker: A deep learning-based system for multiclass vulnerability detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2224–2236, 2021.

[13] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, Z. Deng, and Y. Zhong, "VulDeePecker: A deep learning-based system for vulnerability detection," in *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018. [Online]. Available: https://doi.org/10.14722%2Fndss.2018.23158

[14] W. Huang, S. Lin, and C. Li, "Bbvd: A bert-based method for vulnerability detection," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, 2022. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.01312103

[15] N. S. Harzevili, J. Shin, J. Wang, S. Wang, and N. Nagappan, "Characterizing and understanding software security vulnerabilities in machine learning libraries," in *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, 2023, pp. 27–38.

[16] K. Filus and J. Domańska, "Software vulnerabilities in tensorflow-based deep learning applications," *Computers Security*, vol. 124, p. 102948, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404822003406

[17] Snyk. (2024) Snyk: Developer-first security. Último acceso: 4 de enero de 2025. [Online]. Available: https://snyk.io

[18] L. Leite, C. Rocha, F. Kon, D. Milojicic, and P. Meirelles, "A survey of devops concepts and challenges," *ACM Comput. Surv.*, vol. 52, no. 6, Nov. 2019. [Online]. Available: https://doi.org/10.1145/3359981

RITA: Metodología Para Evaluación de **R**espuestas **I**nteligentes Aplicando **T**axonomías de **A**lucinaciones

Laura Vigoya FIDESOL lvigoya@fidesol.org,
Celia Fernández Rosales FIDESOL cfernandez@fidesol.org
Carmen Alba Jiménez FIDESOL cmjimenez@fidesol.org
Rafael Román FIDESOL rraguilar@fidesol.org
Rubén Morillas FIDESOL rmorillas@fidesol.org

***Resumen*—La inteligencia artificial ha transformado el ámbito tecnológico por su capacidad para generar texto complejo. Uno de los principales desafíos a los que se enfrenta, son las alucinaciones, las cuales constituyen un reto para la ciberseguridad de los usuarios. Este artículo presenta RITA, una metodología para la evaluación de las respuestas generadas por modelos de inteligencia artificial. Su principal aporte radica en incorporar el uso de la taxonomía de alucinaciones para clasificar y categorizar estos fenómenos de manera sistemática. La metodología se validó comparando GPT-4o-mini, Gemini y DeepSeek. GPT-4o-mini destacó como el más consistente con tasas de alucinaciones inferiores al 39 %. La alucinación más recurrente es por lógica inconsistente pues omite información relevante al no jerarquizarla adecuadamente. Esta metodología garantizó la validez y confiabilidad de los resultados, consolidándose como una herramienta clave para la toma de decisiones informadas en el uso de modelos generativos.**

***Index Terms*—Ciberseguridad, LLM, Alucinaciones, detección de alucinaciones, Gemini, DeepSeek, GPT, metodología**

Tipo de contribución: *Investigación original*

I. Introducción

En los últimos años, los modelos de inteligencia artificial (IA) generativa, y en particular los modelos grandes de lenguaje (*Large Language Models*, LLMs), han revolucionado el panorama tecnológico debido a su capacidad para generar texto coherente, responder preguntas complejas y asistir en tareas especializadas [1]. El texto que generan esta basado en probabilidades calculadas a partir de patrones aprendidos y esto hace que no siempre puedan capturar la calidad semántica o la veracidad de las respuestas, lo que conlleva a la aparición de información incorrecta ó **alucinación**. La naturaleza dinámica de la difusión de información en la era digital magnifica estos riesgos, lo que hace necesario el desarrollo de mecanismos sólidos de verificación de hechos, directrices éticas para la generación de contenidos y prácticas de implementación responsables para los LLM. [2].

La evaluación humana (Métrica Cualitativa) permite analizar la coherencia, precisión factual, contexto y la calidad general de las respuestas, detectando errores que los algoritmos no siempre pueden captar. La taxonomía de las alucinaciones permite realizar una clasificación sistemática y organizada de los diferentes tipos de alucinaciones, categorizar y comprender las diversas formas en que los modelos pueden producir información incorrecta en los datos de entrada.

Con el objetivo de abordar el desafío de la desinformación y los errores asociados a la IA generativa, este proyecto propone a **RITA**, una metodología para la evaluación de las **R**espuestas generadas por modelos de **I**nteligencia artificial, considerando su clasificación en función de la **T**axonomía de las **A**lucinaciones. De esta manera, se busca no solo identificar y mitigar las alucinaciones, sino también establecer un marco de evaluación sólido que contribuya a mejorar la fiabilidad de los LLMs en aplicaciones críticas. RITA forma parte del proyecto CICERO, que es el único proyecto de ciberseguridad reconocido como Red de excelencia CERVERA.

II. Revisión de la literatura

La evaluación de los LLMs depende en gran medida de la percepción humana, específicamente de cómo las personas perciben y evalúan sus respuestas, considerando aspectos como los sesgos cognitivos, la estética y la usabilidad. Esto hace necesario un enfoque multidisciplinario que integre áreas como la experiencia del usuario, la educación y la psicología del comportamiento, con el fin de garantizar resultados confiables y alineados con lo que el evaluador espera.

Incorporando detección de alucinaciones en sistematización de procesos humanos en respuestas generadas por la IA generativa, Tam et al. [3], proponen el benchmark FIB (Factual Inconsistency Benchmark) para detectar alucinaciones, comparando resúmenes factualmente coherentes e inconsistentes. TruthfulQA [4] propone un benchmark para validar la factualidad en cuestionarios, ajustando la temperatura del modelo a cero, pero sin especificar una metodología clara. Sin embargo, ninguno de los trabajos encontrados considera la taxonomía de las alucinaciones, ni sus alcances en la evaluación para la comprensión de textos escritos por la IA.

III. Taxonomía de las alucinaciones

Definiendo la alucinación como *contenido generado sin sentido o infiel al contenido fuente proporcionado*[5], La taxonomía de las alucinaciones constituye un sistema de clasificación metodológico y estructurado para categorizar los distintos tipos de alucinaciones generadas por los LLMs.

La metodología presentada en esta investigación realiza una adaptación a La taxonomía presentada por Lei Huan et al.[6], la cual se describe en la figura 1, ajustándola a su uso en la comprensión de textos realizado por la IA.

IV. Metodología propuesta

La evaluación del sistema, tiene principalmente una función diagnóstica. Para que la evaluación del sistema logre su verdadero objetivo y resulte un instrumento útil, debe ejecutarse

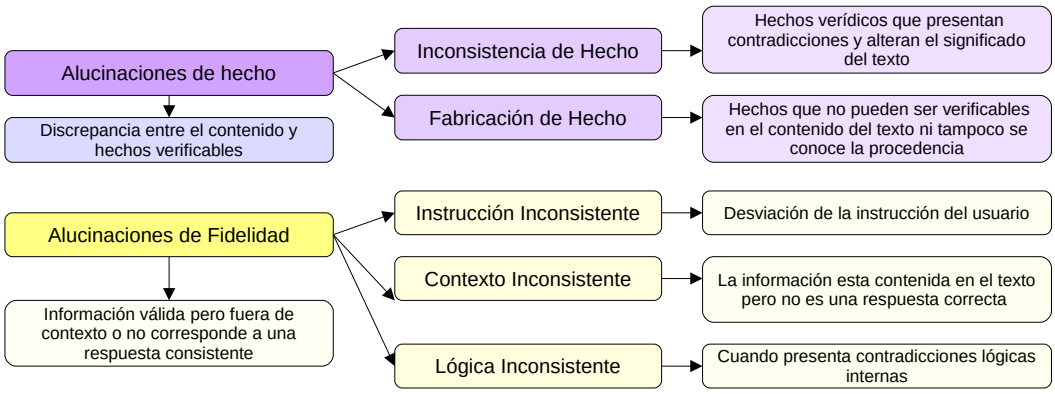


Figura 1. Taxonomia de las alucinaciones.

según un diseño planificado y riguroso desde el punto de vista técnico, garantizando que la información obtenida sea realmente valiosa para los fines que persigue.

IV-A. Descripción de la metodología

Un campo interdisciplinario en donde converjan modelos computacionales de IA y técnicas experimentales de psicología y educación, permite integrar teorías precisas y verificadas sobre el funcionamiento y como aprende la mente humana procesos cognitivos para la IA. Por otra parte, desde hace décadas, España ha implementado diversas iniciativas para evaluar y mejorar las competencias lingüísticas de su población presentes en todos los niveles de enseñanza obligatoria y postobligatoria. Estas evaluaciones se centran en diversas competencias cognitivas específicas, estableciendo un conjunto de pruebas diseñadas para medir el desarrollo de las competencias clave en el alumnado. RITA se consolida siguiendo los lineamientos presentados en el Marco General de las Evaluaciones del Sistema Educativo Español[7], [8], documentos técnicos que proporcionan una base teórica sólida y directrices concretas para la elaboración de instrumentos de evaluación. Basado específicamente en la comprensión lectora, como modelo de evaluación de comprensión de textos para la IA.

IV-B. Conceptos clave

A continuación, se definen y describen una serie de conceptos que proporcionan el marco teórico para la comprensión y aplicación de la metodología diseñada.

IV-B1. Competencia: El concepto de competencia, aplicado a este campo comprende como puede "aprender" y "desempeñarse" la inteligencia artificial en tareas específicas. Aunque la IA no tiene conciencia ni valores éticos propios, puede ser diseñada para imitar ciertas competencias humanas.

IV-B2. Proceso cognitivo: Al igual que para la mente humana, un proceso cognitivo se define aquí como la serie de operaciones que realiza el modelo para procesar información, comprender el entorno, resolver problemas, tomar decisiones y aprender.

IV-B3. Temperatura: En LLM la temperatura es un hiperparámetro que regula la forma de la distribución de probabilidad. Valores bajos (≈ 0) estrechan la distribución, y valores altos (≥ 1) aplanan la distribución, permitiendo una mayor exploración de tokens menos probables. En términos coloquiales, el parámetro de temperatura de un LLM regula la cantidad de aleatoriedad del texto generado, lo que conduce a resultados más o menos diversos, por lo tanto, a menudo se afirma que es el parámetro de creatividad[9].

IV-C. Indicador

El indicador permite precisar las acciones concretas esperadas del proceso cognitivo que se quiere evaluar en la IA. por ejemplo, *"Localizar y obtener información"* para preguntas textuales ó *"Integrar e interpretar ideas e informaciones"* para preguntas de construcción.

IV-D. ítem

En muchas ocasiones la intención de emplear IA es facilitar el proceso de adquisición de la información, es importante identificar el tipo de respuesta. A los tipos de respuesta los denominamos *"Ítems"*. Para el diseño existen tres tipos de ítems:

1. **Respuesta dicotomíca**. Es cuando se forza al modelo a generar una respuesta a la cual debe responder eligiendo entre dos alternativas: verdadero/falso, sí/no, causa/consecuencia, etc.
2. **Respuesta textual**. Este tipo de ítems no requieren de una elaboración o construcción por parte del modelo en cuanto a contenido, sino que solamente busca identificar el sentido literal del texto. Se busca que el modelo no adicione información a la solicitada.
3. **Respuesta semiconstruida**, Este tipo de ítems exige una parte de elaboración de la respuesta por el modelo. En general, solicitan una breve construcción. Puede añadir conocimiento externo o conocimiento previo del texto base a la respuesta de identificación. En estos casos la evaluación de la alucinación debe ajustarse a la asociación con la parte correspondiente a la respuesta de identificación o de elaboración respectivamente.
4. **Respuesta construida**, Este tipo de ítems exige construcción por parte del modelo. Dentro de este tipo

se engloban las producciones escritas que requieren de conocimientos externos al texto base, pero que se alinean con el contenido, sin alterarlo y provienen de las inferencias hechas correctamente por el modelo.

IV-E. Etapas en la metodología

La metodología a seguir para generar el proceso de evaluación consta de los siguientes pasos:

1. **Definición del constructo**: Es la identificación del contexto para el cual necesito el cuestionario y los elementos que se quieren evaluar dentro del texto. Aspectos que interesa medir de acuerdo con un marco teórico
2. **Operacionalización**: Permite identificar las dimensiones y subdimensiones, es decir traduce el o los constructos en un listado jerarquizado y ordenado de indicadores.
3. **Documento base para la construcción del cuestionario**: guiará el trabajo de los responsables de la elaboración de las preguntas que se incorporarán en el cuestionario. Aquí se genera también la elaboración de la codificación de los ítems.
4. **Elaboración del cuestionario**: Diseño y generación del cuestionario de evaluación: Generación de preguntas basándose en los tipos de preguntas. Una vez se han seleccionado las preguntas que se incorporarán al cuestionario, se realiza un pilotaje que permitirá descartar o sustituir aquellas que no informan adecuadamente y se elaborará el instrumento definitivo.
5. **Elaboración de respuestas base**: este servirá de guía para la implementación de la rúbrica.
6. **Elaboración de rúbrica de evaluación**: permite definir los criterios de evaluación y generar la hoja de codificación para cuantificar el impacto de la alucinación.
7. **Aplicación del cuestionario**: permite transformar conceptos abstractos en datos medibles y analizables.
8. **Evaluación de alucinaciones**: Permite sustentar la toma de decisiones a través de juicios elaborados.
9. **Informe de resultados**: Se elabora el documento con los resultados encontrados durante el proceso de evaluación

En su diseño, estas pruebas deben tener en cuenta no sólo los contenidos del criterio evaluado, sino también un formato adecuado a su aplicación y codificación, así como el contexto en el que se emplean.

IV-F. Diseño y elaboración de cuestionarios

Una vez identificados los indicadores, se procede al diseño del cuestionario. A la hora de generar preguntas, las preguntas textuales requieren de exactitud o precisión en la respuesta, lo que las hace más fácil de evaluar pero mas propensas a alucinaciones. Las preguntas semiconstruidas ó construidas permiten al modelo entregar información de su propia producción y son particularmente útiles cuando queremos información detallada o creación en la redacción. Su mayor desventaja es que son más difíciles de codificar, clasificar y verificar para el análisis.

Independientemente de las preguntas que se formulen, hay una serie de características que deben cubrirse al plantearlas:

- Las preguntas tienen que ser claras, precisas y comprensibles para el modelo.
- Es aconsejable que las preguntas sean lo más breves posible, porque las preguntas largas toman más tiempo y son más difíciles de evaluar. La directriz es que se incluyan las palabras necesarias para que se comprenda la pregunta, sin ser repetitivos o barrocos.
- Deben formularse con un vocabulario simple, directo y contextualizado para el modelo.
- Para una correcta codificación. Las preguntas deben referirse preferentemente a un solo aspecto o una relación lógica.

IV-G. Diseño y elaboración de la rúbrica

El uso de la rúbrica se asocia comúnmente con la evaluación del proceso de aprendizaje [10], que permiten al evaluador realizar un seguimiento de una manera más sistematizada. Los LLMs, excepto que se especifique, suelen generar respuestas de elaboración no solicitada, esto ocurre dado que por defecto el modelo está entrenado para proporcionar más información de la requerida, ya sea para aclarar ó contextualizar incluso si la pregunta es textual.

La evaluación supone objetivar algunos criterios de codificación. No obstante, para que la evaluación sea más objetiva y precisa, es aconsejable adaptar la rúbrica a las tareas concretas que formen parte de cada prueba. Recordar que los criterios de evaluación aquí presentados están enfocados a la evaluación de generación de alucinaciones de LLMs en la competencia particular de comprensión de textos.

IV-G1. Codificación de los ítems: Los ítems se codificarán en función al tipo de alucinación que presenten en la taxonomía según la descripción de la Tabla I y se codificarán de acuerdo a las siglas definidas en esta. Es posible que se presente más de una alucinación por ítem y requiera asignarse un valor numérico a cada fallo, categorización de alucinaciones. Esta debe establecerse inicialmente en la rúbrica, de acuerdo a la gravedad de las alucinaciones en el contexto.

IV-H. Análisis de resultados

Las preguntas se plantearán de forma guiada, es decir, conducir desde la orientación de la pregunta, el contexto y los requerimientos esperados, de manera que las producciones resulten lo más exactas posibles a la respuesta planteada en el diseño de la prueba. Partiendo de la base de que los procesos cognitivos están implícitos en la creación del texto se valorara únicamente, la adecuación (la adaptación del texto al contexto), la fiabilidad (veracidad y verificabilidad de los datos) y credibilidad del texto (autoridad y reputación de la fuente). No se harán consideraciones desde el punto de vista lingüístico la coherencia, la cohesión, o el léxico del texto, siempre y cuando no constituyan una alucinación de cualquier tipo. La medición conjunta de la gravedad de la alucinación, puede depender de la agrupación de indicadores relacionados entre sí, lo que permite generar otras escalas de cuantificación para análisis por subdimensiones.

V. Validación de la metodología

La validación de una metodología es un proceso crucial, ya que asegura que los métodos utilizados son apropiados, confiables y capaces de producir resultados válidos y consistentes mediante la aplicación práctica.

Tabla I
CODIFICACIÓN DE ALUCINACIONES EN FUNCIÓN DE LA TAXONOMÍA PARA CADA ÍTEM.

Alucinación	Sigla	Descripción
Dicotómica	ADI	Se le hace una pregunta cuya respuesta, al ser cerrada, simplemente la contesta de manera errónea.
Inconsistencia de Hecho	AHIE	Añade información que no se le está pidiendo o no está dentro del texto pero se conoce como cierta.
Fabricación de Hecho	AHFE	Contiene hechos que no pueden ser verificables en el contenido del texto ni tampoco se conoce la procedencia.
Instrucción Inconsistente	AFI	La inconsistencia aquí significa una desalineación no malintencionada de las instrucciones del usuario. La respuesta no corresponde a la pregunta.
Contexto Inconsistente	AFC	Aunque se encuentra en la información del documento base, no corresponde al contexto de la pregunta y por consiguiente no es una respuesta correcta.
Lógica Inconsistente	AFL	La información está contenida en el documento base, corresponde al contexto en el que se le pregunta, pero la respuesta no corresponde a la lógica de la pregunta, tergiversa el sentido de la respuesta quitando la lógica del documento base

V-A. Definición del constructo

Dentro de las convocatorias de financiación a las que FIDESOL presenta propuestas hay diferentes opciones. Con el objetivo de tener una primera prueba de funcionamiento de la metodología, se seleccionaron las convocatorias Cervera para centros tecnológicos, las Cervera generales para empresas y las de ecosistemas de innovación.

V-B. Operacionalización

Para identificar dimensiones, subdimensiones e indicadores, se contactó con el equipo de expertos para formular los conceptos clave en convocatorias de financiación.

V-C. Construcción del cuestionario, elaboración de la rubrica y las respuestas base

Identificados los contenidos a evaluar y el tipo de respuesta que se requerían del modelo, se procedió a la elaboración de los cuestionarios, la rubrica y la plantilla de evaluación según la metodología. Se ajustaron los prompts al tipo de preguntas que se querían obtener, delimitando las posibles acciones de los motores. Se plantearon 13 preguntas en el cuestionario, todas de respuesta textuales, sin elaboración de preguntas de respuesta dicotómica y considerando que una respuesta se dará como fáctica cuando el modelo proporcione la información que se encuentra en la plantilla de evaluación proporcionada por el equipo de conocimiento experto. El cuestionario diseñado corresponde a un documento genérico que puede ser empleado para cualquier convocatoria de financiación.

V-D. Aplicación del cuestionario

Para la aplicación del cuestionario se seleccionaron tres motores de LLMs, Gemini[11], DeepSeek[12] y ChatGPT[13], ecosistemas que permiten un análisis comparativo robusto, cubriendo diferentes fortalezas y debilidades en el panorama actual de los LLM. Los modelos seleccionados para la aplicación fueron GPT-4o-Mini (ChatGPT), 2.0 Flash (Gemini) y DeepSeek-V3 (DeepSeek).

VI. EVALUACIÓN DE ALUCINACIONES

La detección de las alucinaciones se realizó en dos fases. La primera de ellas consistió en la identificación de las respuestas en fácticas o alucinadas (clasificación binaria), para posteriormente realizar la clasificación de las alucinaciones detectadas dentro de las categorías de alucinaciones propuestas.

VI-A. Detección de alucinaciones

Los resultados obtenidos de la detección de alucinaciones en cada uno de los modelos se presenta en la *Tabla II*.

Tabla II
CANTIDAD DE RESPUESTAS ALUCINADAS EN CADA LLM

Respuesta	DeepSeek	GPT40-mini	Gemini
Fáctica	7	8	4
Alucinada	6	5	9

Tras la aplicación de 13 preguntas a cada uno de los motores, se obtuvo que Gemini entrega el $69,23\,\%$ de las respuestas con algún tipo de alucinación, mientras Gpt-4o-mini solo el $38,48\,\%$ y el $45,15\,\%$ para DeepSeek. Estos resultados permiten establecer a GPT-4o-mini como el modelo con la menor incidencia de alucinaciones entre los evaluados, destacando su mayor precisión en la generación de respuestas.

VI-B. Clasificación de las Alucinaciones

Una vez identificadas las alucinaciones, estas fueron clasificadas de acuerdo a las categorías establecidas, véase *Tabla I*. La *Fig. 2* presenta el porcentaje de alucinaciones en cada una de las categorías. Las preguntas generadas correspondían a textuales, por lo que en la formulación de la pregunta se ajustó la temperatura a cero, excluyendo la presencia de alucinaciones de hecho. Por otra parte, la mayor parte de las alucinaciones generadas por los modelos corresponde a las de alucinación de fidelidad por lógica inconsistente (AFL),pues los motores no son capaces de jerarquizar la información y por lo tanto omiten información relevante de los procesos o escenarios presentados.

Gemini, fue el que presentó una mayor variedad de alucinaciones, y tiene la limitación de no ser capaz de conservar el contexto del documento proporcionado, lo que obligaba a adjuntarlo en cada pregunta. Además, este motor no pudo calcular los tiempos necesarios para procesar y responder cada pregunta. Por su parte, DeepSeek presentó la limitación de sufrir colapso en los servidores durante las pruebas. Esta situación, generaba una perdida de contexto y por consiguiente se hacía necesario volver a entregarle nuevamente el documento de referencia una vez que el servidor volvía a estar disponible, lo que ralentizaba la ejecución de las pruebas. Finalmente, GPT-4o-mini no presentó problemas de ejecución, aunque fue necesario disponer de una cuenta registrada para seleccionar el motor de uso durante la prueba.

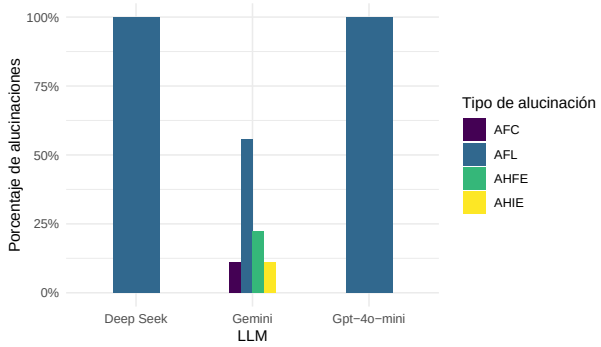


Figura 2. Porcentaje de alucinaciones por tipo y por LLM.

VI-C. Tiempos de ejecución

Durante la ejecución de las pruebas se fueron registrando los tiempos de ejecución de los modelos. El modelo LLM de Gemini no es capaz de registrar los tiempos, y por esto no pudo ser considerado en este análisis. Los cálculos de los estadísticos básicos, sin tener en cuenta las alucinaciones, para cada modelo se presentan en la *Tabla III*

Tabla III
ESTADÍSTICAS BÁSICAS DE TIEMPOS DE EJECUCIÓN DADA POR LOS MODELOS EN SEGUNDOS (S).

	DeepSeek	GPT40-mini
Media	27, 8	12,4
Desviación estándar	50	4
Mínimo	2	8, 6
Máximo	170	22

DeepSeek presenta una notable variabilidad en los tiempos de respuesta, con algunos casos que alcanzan tiempos muy altos, de hasta 170 segundos. La desviación estándar elevada refleja una amplia dispersión en los tiempos de ejecución. Los tiempos no fueron registrados cuando se presentaron fallos en los servidores. En contraste, GPT-4o-mini muestra tiempos de respuesta rápidos y constantes en comparación con DeepSeek. Su baja desviación estándar indica un comportamiento más estable y predecible en lo que a tiempo se refiere, destacándose como el modelo más eficiente en todos los ítems evaluados, en velocidad de ejecución y fiabilidad de los servidores. Sin embargo en las preguntas cortas, cuyas respuestas eran datos muy concretos de fechas o características de entidades, Deep Seek presentaba un mejor rendimiento.

VII. CONCLUSIONES

En este trabajo se ha llevado a cabo el desarrollo, implementación y evaluación de la metodología **RITA**, la cual permite planificar los procesos y aumentar la confianza en los resultados obtenidos. Su relevante contribución consiste en incorporar el uso de una taxonomía de alucinaciones en la evaluación, lo que permite clasificar y categorizar estos fenómenos de manera sistemática.

Como método de evaluación de la metodología, utilizamos la comparación entre los resultados obtenidos por tres diferentes LLMs. En la etapa de detección de las alucinaciones destaca GPT-4o-mini como modelo que menos alucinaciones presenta, seguido por DeepSeek y Gemini, los cuales superan el 45 % de generación de alucinaciones en sus respuestas. En términos de rendimiento, las pruebas revelaron que nuevamente GPT-4o-mini se muestra como el mejor en términos de eficiencia y consistencia.

Dentro de las lineas futuras de investigación es necesario demostrar como la metodología puede ser aplicada por otros investigadores en otros contextos, poblaciones o situaciones y obtener resultados consistentes. Igualmente, planteamos RITA como contribución al campo científico, como herramienta metodológica para la creación de datasets sistematizados para la detección de alucinaciones en LLMs.

AGRADECIMIENTOS

El Proyecto CICERO (Contramedidas Inteligentes de Ciberseguridad para la Red del Futuro) es el único proyecto de ciberseguridad reconocido como Red de Excelencia CERVERA, financiado por el Ministerio de Ciencia e Innovación a través del Centro para el Desarrollo Tecnológico Industrial (CDTI), con cargo a los Presupuestos Generales del Estado 2023 y el Plan de Recuperación, Transformación y Resiliencia. Código de Proyecto: CER-20231019

REFERENCIAS

[1] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 6, Apr. 2024. [Online]. Available: https://doi.org/10.1145/3649506

[2] B. C. Das, M. H. Amini, and Y. Wu, "Security and privacy challenges of large language models: A survey," *ACM Comput. Surv.*, vol. 57, no. 6, Feb. 2025. [Online]. Available: https://doi.org/10.1145/3712001

[3] D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel, "Evaluating the factual consistency of large language models through news summarization," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5220–5255. [Online]. Available: https://aclanthology.org/2023.findings-acl.322/

[4] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," 2022. [Online]. Available: https://arxiv.org/abs/2109.07958

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, p. 1–38, Mar. 2023. [Online]. Available: http://dx.doi.org/10.1145/3571730

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, p. 1–55, Jan. 2025. [Online]. Available: http://dx.doi.org/10.1145/3703155

[7] Ministerio de Educación y Formación Profesional , "Marco general de las evaluaciones del sistema educativo," 2024, .

[8] Ministerio de Educación y Formación Profesional, "Guía para la elaboración de pruebas de evaluación educativa," 2020. [Online]. Available: https://www.libreria.educacion.gob.es/libro/guia-para-la-elaboracion-de-pruebas-de-evaluacion-educativa_169723/edicion/pdf-183202/

[9] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, "Is temperature the creativity parameter of large language models?" 2024. [Online]. Available: https://arxiv.org/abs/2405.00492

[10] S. G. Altamirano Galván, A. L. Méndez Ramírez, and M. B. Rojas Galindo, "Beneficios del uso de la rúbrica en la enseñanza-aprendizaje del diseño," *Zincografía*, Apr. 2022.

[11] G. Team, "Gemini: A family of highly capable multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2312.11805

[12] DeepSeek-AI, "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," 2024. [Online]. Available: https://arxiv.org/abs/2405.04434

[13] OpenAI, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

Sesión Pósters IV: Infraestructuras críticas, IoT y sistemas industriales

A review of: "Standalone vs Hybrid SDN solutions for early detection of network channel attacks in Industrial Control Systems."

Valentine Machaka[1,2] vmachaka@ceit.es
Santiago Figueroa-Lorenzo[2] sfigueroa@unav.es
Saioa Arrizabalaga[1,2] sarrizabalaga@ceit.es
Josune Hernantes[2] jhernates@unav.es

[1]CEIT-Basque Research and Technology Alliance (BRTA), Manuel Lardizabal 15, 20018 Donostia / San Sebastián, Spain.
[2]Universidad de Navarra, Tecnun, Manuel Lardizabal 13, 20018 Donostia / San Sebastián, /Spain.

***Resumen*—This article examines solutions for threat detection and response to attacks on Industrial Control Systems (ICS), through a comparative analysis of standalone and hybrid Software-Defined Networking (SDN) approaches. A virtual cyber-physical testbed was developed following the IEC 62264 Industrial Automation Pyramid (IAP) and the MITRE attack framework for adversary emulation to obtain results. Attack detection utilised packet thresholds, packet analysis (signatures), cryptographic hashing techniques, and attack response through dynamic packet filtering. Research findings reveal that the Standalone SDN is preferable for time-sensitive networks, while the Hybrid SDN suits non-time-sensitive industrial settings. Although Standalone SDN demonstrates a 75 % efficiency improvement, its emerging technology status poses unresolved vulnerabilities and limited testing. Conversely, hybrid SDN ensures robust security and reliability by integrating mature intrusion detection systems (IDS). Thus, choosing the appropriate solution requires weighing performance gains against established security measures.**

***Index Terms*—Network channel attacks, Software-Defined Networking (SDN), Sensor networks, Cybersecurity, Incident detection and response**

Tipo de contribución: *Research already published [1]*

I. Introduction

In a fast-changing world of ICS and Critical Infrastructures (CIs), prioritising network communications' security and reliability is crucial. The growing threat of cyber attacks poses severe risks to these systems, which could result in significant disruptions and operational failures. The water sector has witnessed notable cyber attacks where adversaries gained unauthorised access to their systems. For instance, the 2000 Maroochy Shire Sewage Spill resulted in extensive environmental damage; others, such as the 2021 attempt on the Oldsmar and Pennsylvania water facilities, significantly threatened public health, though unsuccessful, no harmful changes were made to the water supply. These incidents highlight the potential severity of successful cyber intrusions, underlining the urgent need for robust cybersecurity measures to protect essential infrastructure from such threats.

Communication network channel attacks on ICS are categorised into denial of service (DoS-Availability) and Deception attacks (Integrity). Availability attacks deny access to sensor or actuator information, while Integrity attacks involve modifying sensor and actuator information. SDN has emerged as a promising paradigm for enhancing network security and management by providing exceptional flexibility and programmability to enable dynamic threat response and efficient network traffic management by decoupling the control plane from the data plane, allowing for a unified, centralised network perspective [2].

This paper presents standalone and hybrid SDN solutions for the early detection of network channel attacks in the context of wastewater treatment plants (WWTPs). The solutions demonstrate detection and response mechanisms using the SDN architecture capable of real-time assessment of system behaviour and network traffic, indicating security incidents, and providing a real-time response to threats.

II. Related Work

The literature review highlights key trends in related research on detection mechanisms in SDN environments. Among the ten (10) closely related reviewed studies, half reported detection times between 2 and 10 seconds, with 70 % of those utilising hybrid SDN solutions, while the remaining 30 % relied on standalone SDN architectures. The findings also indicate that availability attacks were more prevalent than integrity attacks, underscoring a critical focus area for security research. Additionally, most studies employed software-based simulations such as Mininet, COOJA, and MiniCPS, which, while adequate for general network and system simulations, do not provide realistic ICS environments [3]. The focus of related works was more on attack detection than response. Hence, a gap existed to address these limitations by adopting the MITRE ICS framework, integrating a more comprehensive ICS testbed to enhance the realism and effectiveness of security evaluations, comparing standalone and hybrid solutions, and obtaining standardised metrics for detection and response.

III. Implementation

The virtual cyberphysical testbed was developed in GNS3, encompassing standalone and hybrid SDN scenarios by performing attacks on a simulated WWTP. In the standalone SDN configuration, the RYU controller detects and responds to attacks in the ICS environment. Algorithms are deployed as applications in the northbound interface, and reactive rules are forwarded through the southbound interface. OpenFlow switches are installed in passive-inline mode, enabling visibility and control of traffic when malicious packets are detected. The OpenFlow switch adapts its behaviour based on the perceived threat level. At low threat, it operates passively, transparently forwarding traffic while monitoring it

Standalone SDN solution performance metrics

Attack type	IDS Detection Time	Northbound Response Time	Southbound Response Time	Mitigation Time
T0813 & T0826	0 ms	10 ms	0.4 ms (±0.3)	3 ms (±2)
T0849 & T0842	0 ms	10 ms	0.4 ms (±0.3)	No downtime
T0806	0 ms	10 ms	0.4 ms (±0.3)	3 seconds (±2)

Hybrid SDN solution performance metrics

Attack Type	IDS Detection Time	Northbound Response Time	Southbound Response Time	Mitigation Time
T0813 & T0826	20 ms (±5)	26 ms (±10)	1 ms (±0.5)	18 s (±8)
T0849 & T0842	3.6 s (±3)	5 ms (±4)	1 ms (±0.5)	22 s (±-10)
T0806	20 ms (±5)	5 ms (±4)	4 ms (±3)	3 s (±2)

Figura 1. Summary of SDN Solutions performance metrics.

for insights. When a threat is detected, the switch transforms into an active defender. It enforces security policies received from a controller, such as dropping malicious packets by IP or MAC address, to mitigate threats and actively protect the network.

Contrary to the standalone SDN, the hybrid SDN combines the traditional IDS Snort as a detection engine using signature-based detection. At the same time, the controller is responsible for forwarding rules to the Southbound interface. The IDS is set up through port mirroring in passive mode, and when an alert is triggered, it sends an alert through a network socket to notify the controller of an ongoing attack. The controller then reacts by pushing rules to the switch. In contrast to Loose SDN, the reactive flow installation method (Strict SDN) is considered suitable as an SDN cybersecurity strategy. This strategy is essential for ICS because of its critical requirements for high availability.

The following MITRE ATT&CK techniques were used: T0806 (Brute Force I/O), T0849 and T0842 (Masquerading, Network Sniffing), T0826 (Loss of Availability), T0832 and T0856 (Manipulation of View, Spoof Reporting Message), and T0813 (Denial of Control). Two types of network channel attacks were implemented: availability and integrity. To demonstrate their impact on system availability, three D/DoS attacks are examined, including the Smurf attack, TCP SYN flood attack, and the Modbus multiple registers DoS attack. Regarding integrity attacks, ARP spoofing and the corruption of sensor and actuator data are deployed through the man-in-the-middle (MITM) attack. These tests were conducted through open-source tools like Metasploit, Hping, Ettercap, and a Python-based injection attack script.

IV. Results & Discussions

Although the hybrid solution has a four-fold longer detection time, using an IDS like Snort provides a comprehensive rule set and easy deployment, but may not be ideal for time-sensitive networks. However, the standalone SDN solution performance may come at a potential cost of neglecting mature security technology, potentially exposing critical systems. The results also demonstrate the superiority of a passive-inline device over a passive port mirroring device, attributable to the faster packet transmission facilitated by OpenFlow. Compared to the detection times reported in reviewed studies, the Standalone SDN solution achieves a detection time of 10 ms. In comparison, the Hybrid solution demonstrates a response time of 44 ms, considerably faster than those reported in reviewed studies, which observed detection times ranging between 2–10 s. A comprehensive summary of the detection and response metrics obtained from the proposed methods is presented in Figure 1. Response times varied from 500 μs to 4 ms, while mitigation times ranged from 2 s to 18 s. The hybrid SDN solution is slower as it needs to combine alerting, northbound processing, and southbound rules forwarding time.

V. Conclusions

This research examined detection and response strategies for ICS network channel attacks, comparing standalone and hybrid SDN-based approaches. Results showed that standalone SDN provided faster detection and response times, making it ideal for time-sensitive ICS. However, the hybrid SDN approach offered stronger security. The study highlights the trade-off between speed and security, emphasizing the need for industry-specific solutions. Automated threat detection and response in ICS environments is possible by combining SDN and cybersecurity tools and technologies. Future research will explore machine learning-driven SDN applications to enhance dynamic threat detection. Additionally, anomaly detection in ICS sensor and actuator data will be investigated to mitigate data corruption risks. Lastly, scalability and secure communication between field devices and controllers remain critical for deploying resilient ICS security frameworks.

Acknowledgement

This research was funded by the Elkartek programme of the Basque Government through the project “BEACON: cyBErsecure industriAl Computing cONtinuum” with grant number KK-2023/00085 and the “Ayudas Cervera para Centros Tecnológicos 2023” programme of CDTI (Centre for the Development of Industrial Technology) under the Ministry of Science and Innovation, within the research project “CICERO – Contramedidas inteligentes de ciberseguridad para la red del futuro” with grant number CER-20231019.

Referencias

[1] Machaka, V., Figueroa-Lorenzo, S., Arrizabalaga, S. & Hernantes, J. Comparative analysis of the standalone and Hybrid SDN solutions for early detection of network channel attacks in Industrial Control Systems: A WWTP case study. *Internet Of Things*. **28** pp. 101413 (2024), https://www.sciencedirect.com/science/article/pii/S2542660524003548

[2] Himanshu, Saha, K., Das, P. & De, S. A Network Segmentation Architecture for Flow Aggregation and DDoS Mitigation in SDN Using RAPID Flow Rules. *Proceedings Of The 25th International Conference On Distributed Computing And Networking*. pp. 232-241 (2024), https://doi.org/10.1145/3631461.3631561

[3] Gomez, J., Kfoury, E., Crichigno, J. & Srivastava, G. A survey on network simulators, emulators, and testbeds used for research and education. *Computer Networks*. **237** pp. 110054 (2023), https://www.sciencedirect.com/science/article/pii/S1389128623004991

A Review of "A TabPFN-based intrusion detection system for the industrial internet of things"

Sergio Ruiz-Villafranca* [1], José Roldán-Gómez† [2], Juan Manuel Castelo Gómez† [3], Javier Carrillo-Mondéjar† [2], José Luis Martínez† [1]

[1]Universidad de Castilla-La Mancha, [2]Universidad de Zaragoza, [3]Universidad Politecnica de Madrid,
{luism.garcia, sergio.rvillafranca, joseluis.martinez}@uclm.es, {jroldan, jcarrillo}@unizar.es,
juanmanuel.castelo@upm.es
* Corresponding author † Contributing author

***Abstract*—The rapid expansion of the Industrial Internet of Things (IIoT) has unfortunately led to a surge in security threats targeting its devices and networks. However, deploying effective countermeasures is challenging due to the inherent limitations of IIoT devices, such as constrained processing power and bandwidth, rendering traditional security solutions inadequate. Machine learning-based intrusion detection systems (IDSs) offer a promising avenue, deployable within industrial control systems or at the network edge. A key obstacle for these IDSs is often the scarcity of sufficient traffic data for robust model training in specific industrial settings. This research tackles this challenge by presenting a novel IDS leveraging the TabPFN model, designed to function effectively with limited IIoT traffic data. Its performance was evaluated against other machine learning algorithms (Random Forest, Xgboost, and LightGBM) across varying training dataset sizes and classification complexities. Notably, TabPFN consistently outperformed the others, achieving the most promising results with a 10-20% improvement in key metrics, peaking at an F1 score of 81% for 6-class and 72% for 10-class classification using 1000 training samples.**

***Index Terms*—Industrial Internet of Things, Machine Learning, TabPFN, Cybersecurity**

Tipo de contribución: Investigación ya publicada en "A TabPFN-based intrusion detection system for the industrial internet of things". En The Journal of Supercomputing [1]

I. Extended Abstract

The landscape of industrial operations has dramatically evolved since its initial stages. The current era, often termed Industry 4.0 or the Industrial Internet of Things (IIoT), is marked by the convergence of cutting-edge technologies. These include the Internet of Things (IoT), the analysis of Big Data, high-speed 5G networks, and innovative applications such as artificial vision and advanced Machine Learning (ML) and Deep Learning (DL) techniques aimed at optimizing industrial processes [2]. This integration of modern IT solutions with established Operational Technology (OT) frameworks has been a key driver of this transformation, fostering a synergistic environment that propels the industrial sector towards new horizons of innovation and expansion [3]. Consequently, the incorporation of IIoT technologies into traditional industrial settings introduces a unique set of challenges. These challenges stem from the necessity to uphold the reliability and efficiency of existing industrial workflows while simultaneously adopting novel technologies capable of enhancing them.

The merging of OT and Information Technology (IT) communication standards, alongside the proliferation of IoT and web-based applications, can introduce vulnerabilities if robust security measures are not implemented. These risks are particularly significant for critical infrastructure, where breaches can lead to substantial financial and societal repercussions. Furthermore, many devices deployed within these infrastructures often lack regular security updates [4], [5], making them attractive targets for evolving cyber threats. Therefore, the application of Machine Learning (ML) algorithms in the development of Intrusion Detection Systems (IDS) has emerged as a potent strategy for bolstering the security of IIoT environments [6], [7]. These ML-driven IDSs possess the capability to analyze residual data from malicious activities within networks or applications, providing an effective means for monitoring, prevention, and mitigation of attack impacts [8]. Moreover, the adaptive nature of ML algorithms allows them to learn and respond to novel attack patterns, rendering them particularly valuable against zero-day exploits. This adaptability, combined with their monitoring and response functionalities, significantly reduces the window of opportunity for cyber adversaries [9], [10].

However, the availability of sufficient data for training these sophisticated models can be a limiting factor, contingent on the specific characteristics of the industrial plant or sector. Researchers often encounter challenges in acquiring adequate datasets due to confidentiality concerns and the complexities involved in replicating real-world attack scenarios [11]. To address these data scarcity issues, novel ML approaches, such as the Prior-Data Fitted Network (PFN), have been proposed [12]. These models demonstrate the ability to achieve high performance even when trained on limited datasets, making them well-suited for industrial contexts with stringent data security requirements and inherent device vulnerabilities. This capability arises from the intrinsic design of the TabPFN model, which is optimized for training with up to 1000 data points and classifying up to 10 distinct categories.

This paper introduces an intrusion detection system grounded in the TabPFN model to mitigate these challenges. The primary objective of this system is to provide a framework for the surveillance and detection of anomalies within IIoT environments, utilizing a minimal number of samples during its training phase. Our proposed system is benchmarked against

other leading ML algorithms known for their effectiveness with tabular data [13], [14]. This comparative analysis involved processing and training various ML models on several small datasets, each containing 2500 samples. The aim was to ascertain whether our TabPFN-based approach yields superior results compared to these alternatives. To effectively counter the cybersecurity threats prevalent in IIoT environments, the proposed architecture can be deployed at the network edge connected to the industrial infrastructure or integrated within the IIoT layer's intrusion detection mechanisms, leveraging the industrial control system for its operational functionalities. Considering these aspects, this work contributes in the following key areas:

- **We introduce a novel IDS architecture tailored for intrusion detection in IIoT environments.** This architecture is specifically designed to operate efficiently across diverse IIoT deployments, particularly within emerging industrial network topologies.
- **We develop small-scale IIoT datasets suitable for training and evaluating ML models.** Recognizing the difficulty of obtaining large volumes of real-world data in industrial settings [15], we leverage publicly available IIoT datasets to create smaller, manageable datasets for evaluating our approach with limited training samples.
- **We present a comprehensive performance evaluation of our proposed system against state-of-the-art ML algorithms.** By comparing our TabPFN-based IDS with other established ML techniques using multiple evaluation metrics, we aim to demonstrate its superior classification performance for malicious and benign network traffic.
- **The findings of our study indicate that TabPFN outperforms other analyzed techniques in the majority of tested scenarios.** These results validate the efficacy of our TabPFN-based approach and highlight its potential for deployment as a robust cybersecurity solution within IIoT environments, thereby contributing to the creation of more secure digital industrial ecosystems.

The full version of this paper was published in [1].

Agradecimientos

This work was supported by the Universidad de Castilla-La Mancha under the predoctoral contract 2022-PRED-20677 and the project 2023-GRIN-34056, both financed by the European Social Fund Plus (FSE+), and by the JCCM under the project SBPLY/21/180501/000195. Also, this work is part of the R&D project PID2021-123627OB-C52, funded by the MCIN and the European Regional Development Fund: "a way of making Europe", and it is also funded by MCIN/AEI/10.13039/501100011033 (FEDER, EU) under grant PID2022-142332OA-I00. Javier Carrillo-Mondéjar is also supported by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR under grant TED2021-131115A-I00, by the Spanish National Cybersecurity Institute (INCIBE) under *Proyectos Estratégicos de Ciberseguridad – CIBERSEGURIDAD EINA UNIZAR*, and by the University, Industry and Innovation Department of the Aragonese Government under *Programa de Proyectos Estratégicos de Grupos de Investigación* (DisCo research group, ref. T21-23R).

References

[1] S. Ruiz-Villafranca, J. Roldán-Gómez, J. M. C. Gómez, J. Carrillo-Mondéjar, and J. L. Martinez, "A tabpfn-based intrusion detection system for the industrial internet of things," *The Journal of Supercomputing*, vol. 80, no. 14, pp. 20 080–20 117, Sep 2024. [Online]. Available: https://doi.org/10.1007/s11227-024-06166-x

[2] D. Ivanov, C. Tang, A. Dolgui, D. Battini, and A. Das, "Researchers' perspectives on industry 4.0: multi-disciplinary analysis and opportunities for operations management," *International Journal of Production Research*, pp. 1–24, 08 2020.

[3] P. K. R. Maddikunta, Q.-V. Pham, P. B, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, and M. Liyanage, "Industry 5.0: A survey on enabling technologies and potential applications," *Journal of Industrial Information Integration*, vol. 26, p. 100257, 2022.

[4] L. L. Dhirani, E. Armstrong, and T. Newe, "Industrial iot, cyber threats, and standards landscape: Evaluation and roadmap," *Sensors*, vol. 21, no. 11, p. 3901, 2021.

[5] T. Gueye, Y. Wang, M. Rehman, R. T. Mushtaq, and S. Zahoor, "A novel method to detect cyber-attacks in iot/iiot devices on the modbus protocol using deep learning," *Cluster Computing*, vol. 26, no. 5, pp. 2947–2973, Oct 2023. [Online]. Available: https://doi.org/10.1007/s10586-023-04028-4

[6] B. Babayigit and M. Abubaker, "Industrial internet of things: A review of improvements over traditional scada systems for industrial automation," *IEEE Systems Journal*, vol. 18, no. 1, pp. 120–133, March 2024.

[7] W. Xiang, K. Yu, F. Han, L. Fang, D. He, and Q.-L. Han, "Advanced manufacturing in industry 5.0: A survey of key enabling technologies and future trends," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 2, pp. 1055–1068, Feb 2024.

[8] K. Zidi, K. Ben Abdellafou, A. Aljuhani, O. Taouali, and M. F. Harkat, "Novel intrusion detection system based on a downsized kernel method for cybersecurity in smart agriculture," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108579, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197624007371

[9] M. Al-Ambusaidi, Z. Yinjun, Y. Muhammad, and A. Yahya, "Ml-ids: an efficient ml-enabled intrusion detection system for securing iot networks and applications," *Soft Computing*, vol. 28, pp. 1765–1784, 2023.

[10] Unknown, "Hybrid deep learning enabled intrusion detection in clustered iiot," 2023. [Online]. Available: https://www.techscience.com/cmc/v72n2/47259

[11] X. Cao, Y. Wang, B. Chen, and N. Zeng, "Domain-adaptive intelligence for fault diagnosis based on deep transfer learning from scientific test rigs to industrial applications," *Neural Computing and Applications*, vol. 33, pp. 4483–4499, 2021.

[12] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Meta-learning a real-time tabular automl method for small data," *arXiv e-prints*, pp. arXiv–2207, 2022.

[13] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521002360

[14] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 507–520. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf

[15] L. Li, S. Kumar Damarla, Y. Wang, and B. Huang, "A gaussian mixture model based virtual sample generation approach for small datasets in industrial processes," *Information Sciences*, vol. 581, pp. 262–277, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002002552100935X

A Review of "A dataset to train intrusion detection systems based on machine learning models for electrical substations"

Esteban Damián Gutiérrez Mlot*
CIRCE Centro Tecnológico
Zaragoza, Spain
edgutierrez@fcirce.es

Jose Saldana
CIRCE Centro Tecnológico
Zaragoza, Spain
jmsaldana@fcirce.es

Ricardo J. Rodríguez
University of Zaragoza
Zaragoza, Spain
rjrodriguez@unizar.es

Igor Kotsiuba
Durham University
Durham, UK
igor.kotsiuba@durham.ac.uk

Carlos H. Gañan
Delft University of Technology
Delft, the Netherlands
C.HernandezGanan@tudelft.nl

***Abstract*—This work presents a comprehensive substation traffic dataset to enhance the training and benchmarking of machine learning-based Intrusion Detection Systems. The dataset includes raw network captures and real substation flows, covering standards and protocols such as IEC61850, IEC104, Network Time Protocol, and Precision Time Protocol. It also contains traces of simulated cyberattacks in a controlled environment. The dataset is filtered, anonymized, and it also includes a set of tools for tools for standardized preprocessing, ensuring consistency and reproducibility. This resource addresses the need for high-quality data to optimize IDS in substations, contributing to more secure and reliable electrical distribution networks.**

***Index Terms*—Cybersecurity, Critical infrastructure, Testbed, IEC61850, IEC60870-5-104, IEC104**

Tipo de contribución: *Investigación ya publicada en* "A dataset to train intrusion detection systems based on machine learning models for electrical substations", Data in Brief, vol. 57, p. 111153, 2024, [1].

I. Extended Abstract

Machine learning-based substation IDS can not only identify known threats but also detect new, previously unidentified ones, contributing to safer and more reliable power systems. The effectiveness of these detection systems largely depends on the quality and availability of the datasets used for training. However, the lack of these public datasets [2], [3] limits the training and comparison of different machine learning detection models and IDS implementations. This limitation arises from the difficulty of accessing real infrastructures where industrial protocols and standards are deployed. In this work, thanks to the participation in an international collaborative research project [4], the authors have collected, classified, and processed a significant amount of operational traffic from two real industrial facilities.

The development and maintenance of high-quality datasets is essential for the advancement and deployment of effective IDS solutions. Although this traffic was originally conceived for electrical environments, it is being extended to other critical infrastructure scenarios. We believe that the relevance of the dataset presented in this paper to machine learning-based IDS models is significant, as it improves the accuracy, reliability, and robustness of the models, ensuring comprehensive coverage of potential threats. Furthermore, it facilitates benchmarking and analysis. However, this dataset is primarily intended for use in substations within the power grid.

The dataset provides operational data collected from two substations. The data obtained from the first substation includes frames corresponding to the IEC104 and NTP protocol, while the second one provided data corresponding to the IEC61850 standard and PTP. This dataset also contains attack traces, which where obtained using a testbed with specific hardware in a laboratory setting.

Data has been pre-processed before been ingested by machine learning algorithms. The PCAP files available in the dataset are appropriately filtered and anonymized to prevent the disclosure of sensitive information, such as topology or equipment models, which could be used to attack the critical infrastructure from which the dataset was created. This process is followed by feature extraction, during which CSV files are generated.

Filtering was performed using `tshark` [5]. Due to challenges in handling large files, the files are first split into 10GB chunks, which are then merged after preprocessing. Splitting and filtering were conducted using the *filter_and_split.sh* script, and subsequent merging was performed using the *merge_pcap.sh* script. Both scripts are available in our software repository [1]. Anonymization is carried out using the script *anonymize.sh*, based on `Sanicap` [6].

The final stage of the preprocessing process is feature extraction. Additionally, the dataset provides the original PCAP files to enable users to perform custom feature processing.

The IEC104 protocol operates on top of the transport layer (specifically, over TCP/IP), while the IEC61850 protocol operates on top of the Link layer. This difference requires the use of distinct features for training algorithms. To extract TCP/IP flows relevant to IEC104, we used the `CICFlowMeter` [7] tool. `Tshark` was used to extract features from IEC61850 frames. The dataset includes scripts for feature extraction in each protocol: *generatecsv_iec104.sh* and *generatecsv_iec61850.sh*. A final step in the feature extraction

process is labeling: an additional column, called "Label", is appended to each CSV file to indicate the attack type or the absence thereof, derived from the file name.

An example usage script, *pycaret_ids.py*, is provided to facilitate the execution and comparison of various machine learning algorithms, specifically for classification tasks. This script leverages the `PyCaret` [8] library, an open source tool that simplifies and automates the process of developing machine learning models. The script reads all CSV files from the dataset, categorizes the data using the "Label" column, removes invalid values, and runs several classification models for comparison. It then stores the model with the best results for future predictions.

We ran this script on subsets of the dataset to show how it facilitates model comparison. We employed z-score normalization and StratifiedKFold validation, with a 70% partition for training data. These experiments were run on a machine with two Intel Xeon Gold 2.20GHz and 128GB of RAM. For IEC104, all available traces were used to detect attacks, including DoS, packet flooding, fuzzing, packet starvation, NTP DoS, Port scanning, and PitM. For IEC61850, a single attack was used to illustrate a different type of classification (binary classification).

The results for the IEC104 data indicate that classifier models such as Extra Trees and Random Forest achieve an excellent balance between predictive performance and training time, making them the most suitable ones for real-world applications in this context. Specifically, the Extra Trees classifier achieved the highest accuracy (0.8217) and competitive results in AUC (0.8297), with a moderate training time of 2.620 seconds. Similarly, Random Forest performed well in both AUC (0.9127) and F1-score (0.8059), with a relatively short training time (1.989 s), making it a strong candidate for practical deployment. Results for the detection of fuzzing attacks on the IEC61850 dataset are also presented. LightGBM and Extreme Gradient Boosting provide the best predictive performance (0.8430 and 0.8470 respectively), but incur higher computational costs (186.9080s and 80.1890s respectively). In contrast, Linear Discriminant Analysis offers a solid balance between performance (0.8768 in precision) and efficiency (7.5940s), making it a good choice in situations where fast training is essential. Models such as Ridge Classifier and SVM underperformed (AUC has value 0), while simpler models such as Naive Bayes and K-Neighbors also proved viable.

In conclusion, the dataset presented in this work addresses a need in substation IDS. It allows training and comparing AI models and includes the main protocols and standards used in substation environments (IEC61850, IEC104, NTP, and PTP). Additionally, tools for dataset creation are provided, including filtering and anonymization functionalities. These tools help create new datasets that comply with a standard, ensuring consistency and maintaining data privacy and integrity. Furthermore, the dataset has been extensively tested with `PyCaret`, facilitating the comparison of different classification algorithms, confirming its suitability and reliability for IDS training.

While this dataset represents a valuable step toward standardized benchmarking, it does exhibit certain limitations that leave room for further work. In particular, feature selection remains an ongoing effort. By editing the appropriate configuration files and subsequently rerunning the analysis script, different field combinations can be tested to identify the optimal balance of speed and precision. These experiments can guide future refinements and ensure that the dataset evolves in line with emerging performance requirements.

For the sake of open science and reproducibility, the dataset, along with examples and usage guidelines, is openly distributed under the GNU GPLv3 license. This allows users to modify and adapt it to their needs, promoting collaborative development and innovation.

The full version of this paper was published in [1] **and dataset at** [9].

ACKNOWLEDGMENTS

The research of E. D. Gutiérrez and J. Saldana has been supported by the European Union's Horizon Europe Energy Research and Innovation programme eFORT, Grant Agreement No 101075665. The research of R. J. Rodríguez was supported part by grant PID2023-151467OA-I00 (CRAPER), funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, by grant TED2021-131115A-I00 (MIMFA), funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, by grant *Ayudas para la recualificación del sistema universitario español 2021-2023*, funded by the European Union NextGenerationEU/PRTR, the Spanish Ministry of Universities, and the University of Zaragoza, by grant *Proyecto Estratégico Ciberseguridad EINA UNIZAR*, funded by the Spanish National Cybersecurity Institute (INCIBE) and the European Union NextGenerationEU/PRTR, by grant *Programa de Proyectos Estratégicos de Grupos de Investigación* (DisCo research group, ref. T21-23R), funded by the University, Industry and Innovation Department of the Aragonese Government. The research of C. H. Gañán by the RAPID project (Grant No. CS.007) financed by the Dutch Research Council (NWO).

REFERENCES

[1] E. D. Gutiérrez Mlot, J. Saldana, R. J. Rodríguez, I. Kotsiuba, and C. Gañán, "A dataset to train intrusion detection systems based on machine learning models for electrical substations," *Data in Brief*, vol. 57, p. 111153, 2024.

[2] S. A. Gutierrez, J. F. Botero, N. G. Gomez, L. A. Fletscher, and A. Leal, "Next-Generation Power Substation Communication Networks: IEC 61850 Meets Programmable Networks," *IEEE Power and Energy Magazine*, vol. 21, no. 5, pp. 58–67, Sep. 2023.

[3] S. E. Quincozes, C. Albuquerque, D. Passos, and D. Mosse, "A Survey on Intrusion Detection and Prevention Systems in Digital Substations," *Computer Networks*, vol. 184, p. 107679, 2021.

[4] E. U. H. E. Research and I. programme, "eFORT," [Online; https://efort-project.eu/], 2022, [Accessed 06-10-2024].

[5] Wireshark, "Wireshark," [Online; https://www.wireshark.org], 1998, [Accessed: 2024-06-18].

[6] thepacketgeek, "Sanicap," [Online; https://github.com/thepacketgeek/sanicap], 2014, [Accessed: 2024-06-18].

[7] C. I. for Cybersecurity, "CICFlowMeter - Applications," [Online; https://www.unb.ca/cic/research/applications.html], 2018, [Accessed: 2024-06-19].

[8] M. Ali, "PyCaret: An open source, low-code machine learning library in Python," [Online; https://www.pycaret.org], 2020, [Accessed: 2024-07-15].

[9] E. D. Gutierrez Mlot, "Dataset to Train Intrusion Detection Systems based on Machine Learning Models for Electrical Substations," [Online; https://zenodo.org/records/13898983], 2024, [Accessed 07-10-2024].

Estimación Probabilística del Riesgo en Redes Eléctricas ante Ataques MaDIoT: Un Enfoque Multicapa

Farid Bagheri-Gisour Marandyn, Néstor Rodríguez-Pérez, Mario Castro Ponce, Lukas Sigrist, Gregorio López López
Instituto de Investigación Tecnológica, Universidad Pontificia Comillas, Calle de Alberto Aguilera, 23, 28015 Madrid
{fbagherigisour, nrodriguezp, marioc, lsigrist, gllopez}@comillas.edu

Resumen- **Este artículo presenta una propuesta de modelo para la estimación del nivel de riesgo de una red eléctrica dada a ser afectada por un ataque de tipo MaDIoT (*Manipulation of Demand via IoT*). Para conseguir dicho objetivo, se emplea un diseño multicapa donde la estimación del riesgo global de la red parte de la estimación de riesgo a nivel de dispositivo, a través del estudio de vulnerabilidades conocidas proporcionadas mediante OSINT, seguido de la estimación de riesgo de compromiso de cierta potencia de un nodo de la red, mediante la generación por Monte Carlo de dos funciones de densidad de probabilidad marginales para dos escenarios extremos (*best-case*, *worst-case*), y finalizando por la estimación del riesgo global de la red, a través de la categorización en niveles de la función de densidad de probabilidad conjunta obtenida mediante simulaciones en PowerFactory de diferentes escenarios de nodos para dos modelos de red (IEEE9, IEEE39). Asimismo, el artículo muestra y compara los resultados obtenidos utilizando diferentes algoritmos de *machine learning*, destacando los que proporcionan mejores resultados.**

***Index Terms*- análisis de riesgo, ciberseguridad, redes eléctricas, aprendizaje máquina, *clustering*, *Manipulation of Demand via IoT*.**

Tipo de contribución: *Investigación original*

I. Introducción

Las redes eléctricas modernas enfrentan nuevas amenazas ciberfísicas debido a la proliferación de dispositivos del Internet de las Cosas (IoT) conectados al sistema eléctrico. De entre las amenazas que surgen debido a la implementación de nuevos sistemas IoT, emerge un ejemplo crítico de amenazas que son los denominados ataques MaDIoT (*Manipulation of Demand via IoT*) [1]. En estos ataques, un atacante controla en masa dispositivos IoT de alto consumo (como aires acondicionados, calefactores, cargadores de vehículo eléctrico, etc.) con el objetivo de alterar drásticamente la demanda eléctrica. Estos ataques coordinados pueden provocar efectos tales en la red como inestabilidades de frecuencia y desconexión de generadores [2], sobrecarga de líneas, resonancias y apagones en cascada [1] y/o aumento de los costos operativos [1].

Dado la potencial severidad de los ataques MaDIoT, ha sido crucial desde el mundo de la ciberseguridad, entre otras, la evaluación del riesgo asociado al incremento de dispositivos de IoT (sobre todo de alto consumo) en las diferentes redes eléctricas. Para ello, la literatura aborda este análisis en términos de probabilidad de que ocurra un ataque exitoso [3] y consecuencias esperadas en caso de que ocurra [4-5].

Las metodologías tradicionales de análisis de riesgos se basan en fases predefinidas (identificación de activos, evaluación de amenazas, evaluación de riesgos) y, a menudo, carecen de actualización dinámica, lo que las hace insuficientes para manejar amenazas en constante evolución. La literatura reciente propone metodologías de Evaluación Dinámica de Riesgos (DRA, por sus siglas en inglés), que incorporan fuentes de datos en tiempo real tales como: sistemas de seguridad (IDS/IPS/SIEM) [6-13], datos históricos de vulnerabilidades, como CVSS [12-13], [14-17], conocimiento del sistema, para evaluar continuamente los riesgos [18-19] y conocimiento experto, para la toma de decisiones [20]. Además, los enfoques que aprovechan el aprendizaje automático (por ejemplo, redes bayesianas, redes neuronales) [6-12], [20] y modelos híbridos [21-22] que combinan datos cuantitativos y cualitativos han mostrado resultados prometedores [23].

En este aspecto, la gran mayoría de estudios se centran en metodologías cuantitativas, donde los diferentes sistemas de información y seguridad son utilizados como fuentes principales para poder suministrar de datos a los modelos propuestos. Solo una pequeña parte de los estudios observados en la materia platean sistemas basados en metodologías cualitativas y, por orden de utilización, el último sería los métodos híbridos, provocando que se carezca de propuestas con un mapeo continuo y multicapa desde las vulnerabilidades de cada dispositivo IoT hasta el impacto operativo en la red eléctrica.

Este artículo se centra en el estudio de la probabilidad de que un ataque MaDIoT tenga efecto en la red eléctrica de estudio (IEEE9 e IEEE39) y, asimismo, en la conversión de esa probabilidad en una categoría de riesgo asociado (escalar, acotada), en tiempo real.

Con dicho objetivo en mente, se plantea un modelo multicapa subdividido en 3 tareas principales: estimación del riesgo a nivel de dispositivo, estimación del riesgo de potencia comprometida a nivel de nodo de red y estimación del riesgo global de la red eléctrica bajo estudio.

En referencia a la primera tarea, la estimación del riesgo de los dispositivos se obtiene en función de lo descrito en el artículo [24], donde a través del estudio de las vulnerabilidades asociadas a dichos dispositivos y mediante el enriquecimiento de sus características conocidas (firmware, versión, *vendor*, etc.) por fuentes OSINT como el NIST NVD, se permite conocer el riesgo asociado a ese dispositivo.

Con relación a la segunda tarea, el riesgo de compromiso de cierta potencia de un nodo es obtenida a través de la generación de su función densidad de probabilidad (PDF) marginal de compromiso de potencia. Esta PDF, al no poder ser obtenida empíricamente de antemano, ya que no se conocen los dispositivos a priori, se construye a través de

simulaciones por Monte Carlo de dos escenarios extremos, a los que denominamos *best-case* y *worst-case*. Estos escenarios son implementados mediante dos variables aleatorias que los modelan.

Por lo que se refiere a la tercera tarea, la estimación del riesgo global de la red, se generó una subdivisión de éste en dos flujos: *offline* y *online*.

En la parte *offline* se desarrollan dos tareas: el proceso relativo a la obtención de la PDF conjunta de compromiso de la red de estudio, descrita mediante las cargas deslastradas y/o protecciones activadas y la categorización de esas PDFs conjuntas en niveles de riesgo. Esta categorización en niveles de riesgo se realiza mediante la creación de umbrales (para cada una de las características que describen la PDF conjunta), a través del estudio y selección de diferentes modelos de *clustering*

Por último, en la parte *online*, se desarrolla el flujo encargado de determinar, en tiempo real, esa categoría de riesgo asociado a la red. Para ello, se ingesta las entradas relativos a los dispositivos que posee la red y la propia topología de la red y se pasan a comparar con las simulaciones de los diferentes casos de la parte offline, obteniendo el valor de riesgo asociado.

II. Metodología

El diseño del modelo propuesto (véase Fig. 1) consta de una arquitectura que engloba, a alto nivel, 3 bloques diferenciados (A, B y C) que, aunque trabajen de manera conjunta al formar parte del mismo flujo de datos, realizan tareas diferenciadas unas de otras, las cuales se listan a continuación:

- **Bloque A**: Subproceso encargado de la ingesta de los ficheros relativos a las características de los dispositivos (tipología, producto, marca, versión, geolocalización, CPE y potencia máxima) y a la topología de la red (actualmente relativa a los modelos IEEE9 y IEEE39).
- **Bloque B**: Subproceso encargado del enriquecimiento de los ficheros ingestados en el Bloque A referentes a los dispositivos. Este segundo bloque se realiza según lo descrito en [24].
- **Bloque C**: Subproceso encargado de determinar y categorizar el riesgo global de la red eléctrica. Dicho bloque se subdivide a su vez en dos tareas interrelacionadas y dependientes: estimación del riesgo de compromiso de cierta potencia de un nodo y estimación del riesgo global de la red eléctrica.

Con el objetivo de poder calcular a tiempo real el riesgo asociado a la red (y previamente a la de los nodos particulares), las subtareas del Bloque C se dividen en dos flujos paralelos, a las que denominamos *offline* y *online.*

El flujo *offline* es el encargado de realizar todos los cálculos a priori necesarios para tener una base empírica con la que poder realizar la clasificación del riesgo. Estos cálculos son referentes a:

- Generación de las PDFs marginales de los dos casos extremos (*best-case, worst-case*).
- Utilización de las PDFs marginales para la simulación en PowerFactory de los efectos en red relativos a cargas deslastradas y protecciones activadas. Los datos obtenidos permiten constituir la PDF conjunta de la red.
- Empleo de la PDF conjunta de la red para calcular y establecer umbrales en cada una de las características que la componen.
- Estimación de la categoría de riesgo (escalar, acotado) de la red a partir de los umbrales establecidos anteriormente.



Fig. 1: Arquitectura de estimación y categorización de riesgo propuesta.

En contraposición, el flujo *online* es el encargado de gestionar todo el ciclo de datos en tiempo real, utilizando según corresponda, los valores previos obtenidos en el flujo *offline*, con el objetivo de generar la categoría de riesgo asociada a la red de estudio. Para ello, se definen los siguientes pasos:

- Cálculo de la PDF marginal de los nodos a tiempo real en base a los valores de riesgo de cada uno de los dispositivos que componen el nodo, recogido en el fichero enriquecido del Bloque B.
- Comparación iterativa de la PDF marginal de cada nodo obtenida en el flujo *online* con las dos PDFs marginales de los escenarios *best-case*, *worst-case* definidos y generados en el flujo *offline.*
- Selección en base a la comparación de las PDFs marginales y de la PDF conjunta de la red que corresponda a los escenarios (*best-case*, *worst-case*) seleccionados por cada nodo.
- Adquisición de los umbrales asociados a la PDF conjunta seleccionada, calculadas en el flujo *offline*.
- Obtención de la categoría de riesgo asociada a la red.

En aras de precisar cada uno de los pasos que se realizan en ambas ramas, se detalla en las siguientes secciones los diferentes hitos que se han listado anteriormente.

A. Offline - Generación de las PDFs marginales de los casos escenario de los nodos

Como consecuencia de que no se posee información a priori relativa a la cantidad y características de los dispositivos que conforman la red bajo estudio, se generan dos casos de escenarios extremos con los que poder modelar los nodos: *best-case, worst-case*. El objetivo de estos escenarios es crear dos contextos de ataque diferenciados que pueda sufrir una red: uno en el que el ataque no suponga una gran amenaza (pocos dispositivos, con un riesgo bajo de compromiso y una potencia baja) y otro en el que el ataque sí suponga una amenaza considerable (muchos dispositivos, con un riesgo alto de compromiso y una potencia alta).

Para la conformación de estos escenarios, se realiza la configuración de una variable aleatoria que se caracteriza por medio de tres distribuciones:

- **Cantidad de dispositivos**: Modelado a través de una distribución uniforme $X \sim \mathcal{U}(a,b)$, siendo *Best-case* $X \sim \mathcal{U}(5,100)$ y *Worst-case* $X \sim \mathcal{U}(100,2000)$.
- **Probabilidad de compromiso del dispositivo**: Modelado a través de una distribución $Y \sim \text{Bernoulli}(p)$ donde p queda definido a través de una distribución uniforme $p \sim \mathcal{U}(a,b)$. En el *Best-case*: p $\sim \mathcal{U}(0,0.3)$ y en el *Worst-case*: $p \sim \mathcal{U}(0.7,1)$.
- **Potencia comprometida del dispositivo**: Modelado a través de una distribución uniforme $Z \sim \mathcal{U}(a,b)$ con un rango para el *Best-case* $Z \sim \mathcal{U}(0.001,0.1)$ MW y para el *Worst-case* $Z \sim \mathcal{U}(0.1,0.5)$ MW.

Tras caracterizar la variable aleatoria que modela los nodos en cada uno de los escenarios establecidos, mediante simulaciones por Monte Carlo, se realiza el cálculo de la PDF y posteriormente de la función densidad de probabilidad acumulada (CDF). Las simulaciones Monte Carlo constan de 1.000.000 de iteraciones para caso. Todos los valores fueron normalizados en función de la P_{max} de cada uno de los nodos de los modelos de red eléctrica utilizados.

B. Offline – Generación de la PDF conjunta de la red

El objetivo de esta subtarea es generar la PDF conjunta de la red para que pueda ser caracterizada y poder estimar su riesgo.

Como no se posee información a priori relativa a la topología de la red de estudio (está vendrá definida posteriormente en la rama *online*, si es proporcionada, al igual que los dispositivos), fue necesario definir modelos de redes eléctricas a analizar. Como aproximación para este estudio, se seleccionar los modelos de red IEEE9 e IEEE39, ampliamente utilizados en investigación. Dichos modelos de red poseen, respectivamente, 3 y 19 nodos con demanda conectada.

Utilizando la herramienta de simulación DIgSILENT PowerFactory, por cada modelo de red y por cada tipo de escenario de nodo, se pasó a generar las simulaciones que nos permitan poder generar la PDF conjunta. Para el modelo IEEE9 (3 nodos con demanda) se establecen 1000 simulaciones para cada uno de los casos escenario: *best-case*, *worst-case* y *mix-case* (siendo esta, combinación de las 2 anteriores) y por cada nodo. Análogamente, para el modelo IEEE39 (19 nodos con demanda), se establecen 1000 simulaciones para cada uno de los mismos casos escenario y por cada nodo (excepto *mix-case*, que solo tendrá configuración de 19 nodos debido a la explosión combinatoria).

De los resultados de las simulaciones se utilizan las cargas deslastradas (MW) y las protecciones activas (subfrecuencia, sobrefrecuencia, subtensión, sobretensión), para generar la PDF conjunta, ya que proporcionan información de contingencia ante anomalías en la red (sobrecarga) e interpretación directa con posteriores valores de umbrales.

C. Offline – Cálculo de los umbrales en las características objetivo

Dentro de la metodología aplicada, los umbrales se tienen que definir para poder delimitar rangos de valores dentro de cada una de las características que se seleccionan en la PDF conjunta de la red. A esos rangos de valores, posteriormente se le asociará un nivel de riesgo.

Para poder establecer cuáles son los umbrales que mejor ajusten esos rangos de valores, se utilizaron y compararon diferentes algoritmos de agrupamiento (clustering) para obtener la mejor aproximación. En concreto, los algoritmos de agrupamiento empleados incluyeron K-Means++, GMM y DBSCAN para la generación de los diferentes clústeres.

Los algoritmos de agrupamiento en sí no proporcionan directamente los umbrales entre los clústeres, sino que identifican los clústeres y sus centroides. Para determinar los umbrales dentro de este marco, se calculó el valor medio de las distancias entre los centroides (o medias, en el caso de los algoritmos probabilísticos) de los clústeres ordenados de forma consecutiva para las características seleccionadas (deslastre de cargas, protecciones activas).

D. Offline – Estimación del riesgo de la red en base a los umbrales

La estimación del riesgo de la red estará determinada en función de los valores de las características seleccionadas arrojadas por las simulaciones. Lo que hay que determinar es la asociación de esos valores numéricos con una categoría o nivel de riesgo. Esta asociación se realizó utilizando los umbrales como *proxy*. El proceso, es el siguiente:

1. Por cada una de los modelos de *thresholding*, se utilizan los valores de los *thresholds* para separar las PDFs de las características en rangos.
2. De forma ordenada e incremental, se asocian esos rangos de valores a niveles de riesgo (de menor a mayor riesgo).
3. Un evento, proveniente del flujo online será comparado con las diferentes PDFs de las características de estudio del modelo de red que provenga el evento.
4. La comparación con todas las características provoca la conformación de un vector de Riesgo ($\text{Risk}_{\text{vector}}$)
5. El valor final de riesgo de red será el valor máximo de ese vector de riesgo: $\max(Risk_{vector})$

E. Online – Generación de la PDF marginal de los nodos en tiempo real

Con respecto a la rama *online*, para poder obtener la estimación del riesgo de la red a tiempo real, es necesario calcular las PDFs marginales de cada nodo en función de los datos que se ingesten del Bloque B.

Uno de los métodos matemáticos posibles para calcular una PDF mixta compuesta de múltiples variables aleatorias con sus propias PDFs es utilizar la operación convolución.

Con ello y mediante la utilización de la *Fast Fourier Transform* (FFT), que mejora la eficiencia del cálculo de la convolución, se realiza el cálculo de la PDF marginal en tiempo real por cada uno de los nodos. El proceso de cálculo es el siguiente:

1. Cada uno de los dispositivos se modelan como una variable de Bernoulli, cuyos valores son los relativos a no comprometer potencia (0 MW) o comprometer la potencia máxima del dispositivo ($P_{maxdisp}$). La probabilidad de compromiso se calculó a partir de la valoración de riesgo calculada en el Bloque B.
2. Modelados los dispositivos mediante las distribuciones, se pasa a calcular la PDF marginal del nodo aplicando la FFT. El resultado es la obtención de la PDF del nodo.

El procedimiento es posteriormente aplicado a cada uno de los nodos que conforman la red. El proceso se sintetiza en la Fig. 2

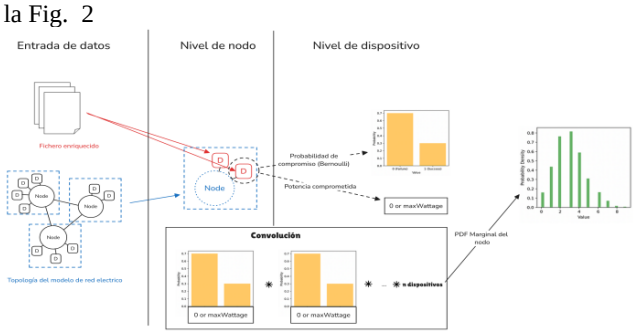


Fig. 2: Online - Esquemático del proceso de generación de la PDF marginal en tiempo real.

F. *Online – Estimación del riesgo global a tiempo real*

La estimación del riesgo global en la rama *online* se apoyará en lo ya realizado en su contraparte *offline*. Al haber obtenido las diferentes PDFs marginales para cada uno de los nodos, se propone es el siguiente proceso (véase Fig. 3):

- Por cada una de las PDFs marginales obtenidas en el flujo *online*, se utiliza la divergencia de Kullback-Leibler para compararlo con los 2 escenarios elaborados en la parte *offline*, seleccionando la más semejante. Este proceso se realiza para todos los nodos.
- Habiendo obtenido a que caso es más semejante por cada nodo, se obtienen todas las simulaciones realizadas para esa configuración, promediando los valores.
- La simulación promediada obtenida es la utilizada para poder identificar, dentro de los rangos de valores generados por los umbrales, a cuál pertenece.
- Conociendo el rango del valor al cual pertenece, se consigue el nivel de riesgo, ya que previamente, en la parte *offline*, ya se había realizado esa asociación.

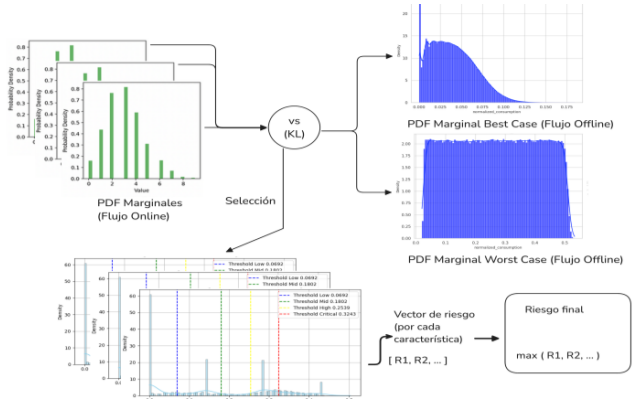


Fig. 3: Online – Esquemático del flujo de estimación riesgo global

III. Resultados

Para evaluar el desempeño de la metodología propuesta, se analizaron los resultados de la estimación del riesgo global en las redes de estudio IEEE9 e IEEE39, considerando los distintos escenarios de ataque.

Inicialmente, se presentan los mapas de calor (Fig. 4 y Fig. 5) que ilustran la asignación de categorías de riesgo (eje X) en función de los diferentes casos (eje Y) para los modelos de thresholding aplicados en el flujo *offline*.

Las categorías de riesgo que se asignan son valores numéricos enteros positivos (empezando por 0). El conjunto de valores de los mapas de calor indica la cantidad de simulaciones realizadas para cada tipo de caso.

La Fig. 4 muestra, la estimación del riesgo que realizan los diferentes modelos para la red IEEE9. En estos mapas de calor se observa que:

- Los casos escenario *best-case* siempre son valorados con el nivel de riesgo más bajo. Esto permite indicar que el aporte de riesgo a la red al ser atacado por este tipo de caso es mínimo.
- Los casos escenario *mix-case* son siempre valorados de igual forma que su contraparte del escenario *worst-case*, en cualquier modelo. Debido a que el aporte en riesgo del *best-case* es mínimo, el aporte en riesgo en la red proviene casi exclusivamente de los nodos que se simularon bajo el escenario *worst-case*.
- Los casos escenario *worst-case* difieren en la estimación del riesgo en función del modelo empleado. El umbral de DBSCAN es más disperso, mientras que los umbrales de K-Means++ y GMM proporcionan una estimación más progresiva.

La Fig. 5 se centra exclusivamente en los escenarios de peor caso para la red IEEE39 debido a las similitudes observadas con los comportamientos del *best-case* y *mix-case* de IEEE9 (Fig. 4). Las observaciones de estos mapas de calor describen que todas las metodologías de umbrales basadas en agrupamiento generan una estimación progresiva y diagonal. Específicamente, el umbral de K-Means++ y el umbral de GMM tienden a la sobreestimación (valores por encima de la diagonal), mientras que el umbral de DBSCAN tiende a la subestimación (valores por debajo de la diagonal).

Con el objetivo de sintetizar los resultados obtenidos y poder escoger la modalidad que mejor permite calcular valores de umbrales, en la Tabla I se desglosan las propiedades observadas por cada una de ellas.

Durante todo el análisis comparativo (véase Fig. 4, Fig. 5, Tabla I) se observó que el algoritmo GMM demostró consistentemente una superior variabilidad intra-caso y dispersión inter-caso, adaptándose eficazmente a las distribuciones de datos reales.

A diferencia de otros métodos de agrupamiento, GMM ofrece transiciones suaves de umbrales entre los niveles de riesgo, lo que da lugar a una categorización más coherente y representativa de los escenarios de ataque MaDIoT, lo que llevó a su selección.

IV. Conclusiones

Este artículo muestra que la metodología multicapa que se propone para la estimación de riesgo global en las redes de estudio IEEE9 y IEEE39 al ser atacadas por ataques MaDIoT es explicativa y consistente con un análisis clásico de riesgo.

Se puede apreciar que tanto la recopilación e inventariado de los dispositivos presentes en el Bloque A, como el análisis de vulnerabilidades y riesgo de dichos dispositivos en el Bloque B son propios de la literatura clásica del análisis de riesgo. Sin embargo, la metodología propuesta presenta novedades en el bloque C, al utilizar métodos probabilísticos y metodologías de *thresholding* para estimar niveles de riesgo. Estas metodologías, aun difiriendo significativamente respecto a metodologías más clásicas, sí que permiten la

misma interpretación con respecto a valores de riesgo provistos empíricamente.

No obstante, es necesario evaluar la metodología presentada con un mayor número de modelos de red diferentes, con el fin de ampliar su validación.

AGRADECIMIENTOS

Este trabajo fue desarrollado en el marco del Proyecto eFORT, proyecto financiado por el programa de Investigación e Innovación Horizonte Europa de la Unión Europea bajo el Grant Agreement No. 101075665.

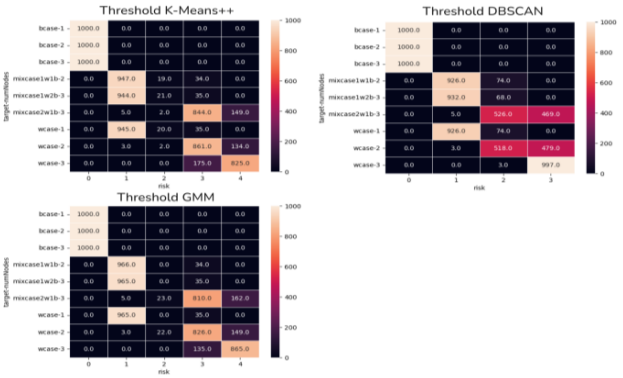


Fig. 4: Mapas de calor por cada modelo de *thresholding* en la red IEEE9 (*best-case*, *mix-case*, *worst-case*).

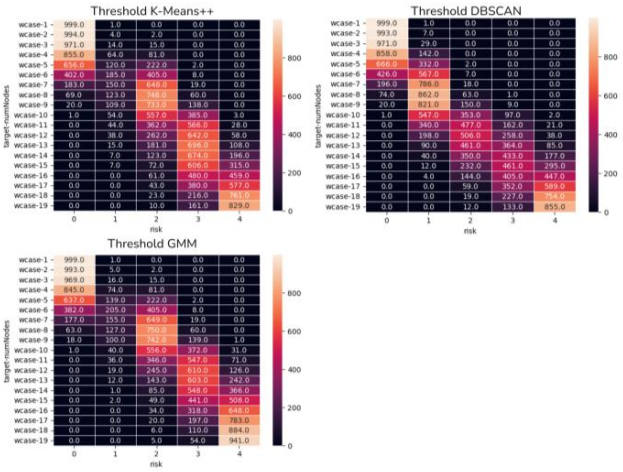


Fig. 5: Mapas de calor por cada modelo de *thresholding* en la red IEEE39 (*worst-case*).

Tabla I
SELECCIÓN DE MODELO DE THRESHOLDING

Modelo	Propiedades
Threshold K-Means++	Mejora en la variabilidad intracaso y dispersión intercaso, pero inferior a GMM. Generación de clústeres esféricos, sin adaptación a la forma real.
Threshold DBSCAN	Mejora en la variabilidad intracaso y dispersión intercaso en ciertas redes, pero inconsistente. Imposibilidad de ajuste en la cantidad de thresholds generados, ya que está basado en otros parámetros. Clasificación de ciertos valores como ruido
Threshold GMM	Variabilidad intracaso y dispersión mejor ajustada intercaso. Mejora en la transición suave intracaso. Permite adaptarse a la forma del clúster.

REFERENCIAS

[1] S. Soltan, P. Mittal, y H. V. Poor, «BlackIoT: IoT Botnet of High Wattage Devices Can Disrupt the Power Grid», 2018.

[2] S. Maleki, S. Pan, S. Lakshminarayana, y C. Konstantinou, «Survey of Load-Altering Attacks Against Power Grids: Attack Impact, Detection and Mitigation», 29 de octubre de 2024.

[3] J. Ospina, X. Liu, C. Konstantinou, y Y. Dvorkin, «On the Feasibility of Load-Changing Attacks in Power Systems During the COVID-19 Pandemic», *IEEE Access*, vol. 9, pp. 2545-2563, 2021.

[4] X. Liu, M. Shahidehpour, Z. Li, X. Liu, Y. Cao, y Z. Li, «Power System Risk Assessment in Cyber Attacks Considering the Role of Protection Systems», *IEEE Trans. Smart Grid*, pp. 1-1, 2016.

[5] N. Rodríguez-Pérez, J. Matanza Domingo, L. Sigrist, J. L. Rueda Torres, and G. López López. Confronting the Threat: Analysis of the Impact of MaDIoT Attacks in Two Power System Models. Energies, 16(23), 7732, 2023.

[6] H. Cam y P. Mouallem, «Mission assurance policy and risk management in cybersecurity», *Environ. Syst. Decis.*, vol. 33, n.º 4, pp. 500-507, dic. 2013.

[7] H. Cam, «Risk assessment by dynamic representation of vulnerability, exploitation, and impact», presentado en SPIE Defense + Security,p. 945809, may 2015.

[8] Q. Zhang, C. Zhou, N. Xiong, Y. Qin, X. Li, y S. Huang, «Multimodel-Based Incident Prediction and Risk Assessment in Dynamic Cybersecurity Protection for Industrial Control Systems», *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 46, n.º 10, pp. 1429-1444, oct. 2016.

[9] Q. Zhang, C. Zhou, Y.-C. Tian, N. Xiong, Y. Qin, y B. Hu, «A Fuzzy Probability Bayesian Network Approach for Dynamic Cybersecurity Risk Assessment in Industrial Control Systems», *IEEE Trans. Ind. Inform.*, vol. 14, n.º 6, pp. 2497-2506, jun. 2018.

[11] Q. Zhu, Y. Zhao, L. Fei, y C. Zhou, «A Dynamic Decision-Making Approach for Cyber-Risk Reduction in Critical Infrastructure», en *IEEE 8th Annual International Conference on CYBER Technology*, jul. 2018.

[12] J. K. Debnath y D. Xie, «CVSS-based Vulnerability and Risk Assessment for High Performance Computing Networks», en *2022 IEEE International Systems Conference (SysCon)*, abr. 2022, pp. 1-8.

[13] I. Kotenko y E. Doynikova, «Security metrics for risk assessment of distributed information systems», en *IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, sep. 2013, pp. 646-650.

[14] K. Huang, C. Zhou, Y.-C. Tian, W. Tu, y Y. Peng, «Application of Bayesian network to data-driven cyber-security risk assessment in SCADA networks», en *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)*, pp. 1-6. nov. 2017,

[15] Y. Qin, Y. Peng, K. Huang, C. Zhou, y Y.-C. Tian, «Association Analysis-Based Cybersecurity Risk Assessment for Industrial Control Systems», *IEEE Syst. J.*, vol. 15, n.º 1, pp. 1423-1432, mar. 2021.

[16] S. Abraham y S. Nair, «A Novel Architecture for Predictive CyberSecurity Using Non-homogenous Markov Models», en *2015 IEEE Trustcom/BigDataSE/ISPA*, ago. 2015, pp. 774-781.

[17] Q. Hong *et al.*, «An information security risk assessment method based on conduct effect and dynamic threat», *8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, nov. 2017, pp. 782-786.

[18] G. Gonzalez-Granadillo *et al.*, «Dynamic risk management response system to handle cyber threats», *Future Gener. Comput. Syst.*, vol. 83, pp. 535-552, jun. 2018.

[19] P. K. Vaddi, Y. Zhao, y C. Smidts, «Dynamic Probabilistic Risk Assessment for Cyber Security Risk Analysis in Nuclear Reactors.», 2022.

[20] «CPS Information Security Risk Evaluation Based on Blockchain and Big Data», *Teh. Vjesn. - Tech. Gaz.*, vol. 25, n.º 6, dic. 2018.

[21] Z. Wang, L. Chen, S. Song, P. X. Cong, y Q. Ruan, «Automatic cyber security risk assessment based on fuzzy fractional ordinary differential equations», *Alex. Eng. J.*, vol. 59, n.º 4, pp. 2725-2731, ago. 2020.

[22] K. Yan, X. Liu, Y. Lu, y F. Qin, «A Cyber-Physical Power System Risk Assessment Model Against Cyberattacks», *IEEE Syst. J.*, vol. 17, n.º 2, pp. 2018-2028, jun. 2023.

[23] G. Gonzalez-Granadillo *et al.*, «Automated Cyber and Privacy Risk Management Toolkit», *Sensors*, vol. 21, n.º 16, Art. n.º 16, ene. 2021.

[24] V. G. Fernández, N. Rodrígºuez Pérez, R. G. Miñarro, J. M. Domingo, R. P. Hielscher, y G. López López, «Dynamic risk assessment tool for customer IoT infrastructures for Smart Grids», en *2023 JNIC Cybersecurity Conference (JNIC)*, jun. 2023, pp. 1-4.

A review of: A Novel Protocol Using Captive Portals for FIDO2 Network Authentication

Martiño Rivera-Dourado* RNASA-IMEDIR group, CITIC research center, Universidade da Coruña A Coruña, Spain
Marcos Gestal RNASA-IMEDIR group, CITIC research center, Universidade da Coruña A Coruña, Spain
Alejandro Pazos RNASA-IMEDIR group, CITIC research center, Universidade da Coruña A Coruña, Spain
Jose Vázquez-Naya RNASA-IMEDIR group, CITIC research center, Universidade da Coruña A Coruña, Spain

Abstract—FIDO2 authentication is increasingly used in web services to replace passwords and mitigate their vulnerabilities. However, its integration with network authentication remains unexplored. This paper summarises FIDO2CAP, a novel protocol for captive-portal authentication using FIDO2 authenticators, including security keys and passkeys. To validate this approach, we developed a prototype and conducted a usability study with 15 users in a mock scenario. Our work represented the first systematic adaptation of network authentication to the FIDO2 paradigm.

Index Terms—WebAuthn, Network Authentication, Captive Portal, Protocol, FIDO2, Security key, Authenticators, Passkey

Type of contribution: Already published research [1]

I. Introduction

Passwords are widely used for authentication due to their simplicity, but they are vulnerable to attacks like phishing and data breaches, which expose users to significant security risks. To address these issues, the FIDO Alliance was established in 2014, introducing the FIDO CTAP standards and security keys like Yubikey and Google Titan. In 2019, the WebAuthn API brought FIDO2 authentication to web applications, leading to its adoption by major browsers. The use of "passkeys", the commercial term for FIDO2 credentials, grew by 50% from 2019 to 2022 [2], with widespread integration across platforms like Google, Microsoft, Apple, and Facebook.

The WebAuthn API [3] defines two main ceremonies: registration (attestation) and authentication (assertion). During registration, a FIDO2 public key and credential identifier are stored on the authentication server. In authentication, the user proves their identity by signing a challenge with their FIDO2 credential's private key. FIDO2 credentials can be either discoverable (client-side, stored on a hardware device) or non-discoverable (server-side, encrypted in the credential identifier). Each credential type has different registration and authentication flows. Our proposed FIDO2CAP protocol supports both credential types to ensure device compatibility.

Although FIDO2 authentication has been used in web applications, it has not yet reached Network Access Control (NAC). NAC systems play a key role in access control, ensuring that only authorized devices can communicate within a network. Corporate wireless networks use protocols like WPA2/3-Enterprise to control access through the Extensible Authentication Protocol (EAP) and IEEE 802.1X at the link level. However, many EAP configurations still depend on passwords for authentication. On the other hand, captive portals are used to restrict access to network resources until the user completes a web-based authentication process. In this context, our novel FIDO2CAP protocol aims to integrate FIDO2 credentials in captive portals though the W3C WebAuthn API.

A. Contribution

In our already published work [1], we introduced the FIDO2 Captive-portal Authentication Protocol (FIDO2CAP), integrating FIDO2 in captive portal network authentication. Besides, we presented a prototype of FIDO2CAP on top of OpenNDS. This system enhances captive portal authentication by making it resistant to common attacks targeting password-based systems. Finally, for validation, we have deployed our prototype on networking hardware, ensuring compatibility with various operating systems and web browsers, and conducted a usability experiment with 15 users.

II. FIDO2CAP protocol

We propose the FIDO2 Captive-Portal Authentication Protocol (FIDO2CAP), which uses FIDO2 for network authentication in captive-portal environments. This section outlines the architecture design and protocol operation.

A. Architecture Design

The FIDO2CAP system is based on the captive-portal framework defined in RFC 8952 [4]. Figure 1 illustrates the architecture, highlighting the proposed modifications in blue. Key components include the FIDO2 Authenticator for user identity verification, a web browser with the WebAuthn API for authentication, the WebAuthn Authentication Web Application (WAWA) as the user portal, a User Database for storing FIDO2 credentials, an optional Session Database for tracking open sessions, and a Registration Portal that supports both administrator and self-registration of FIDO2 Authenticators.

B. Protocol Operation

FIDO2CAP encompasses two main ceremonies: user authentication and FIDO2 registration. While the user registration can be performed outside the network environment via a WebAuthn API web application, FIDO2CAP

ISBN:978-84-10169-61-6

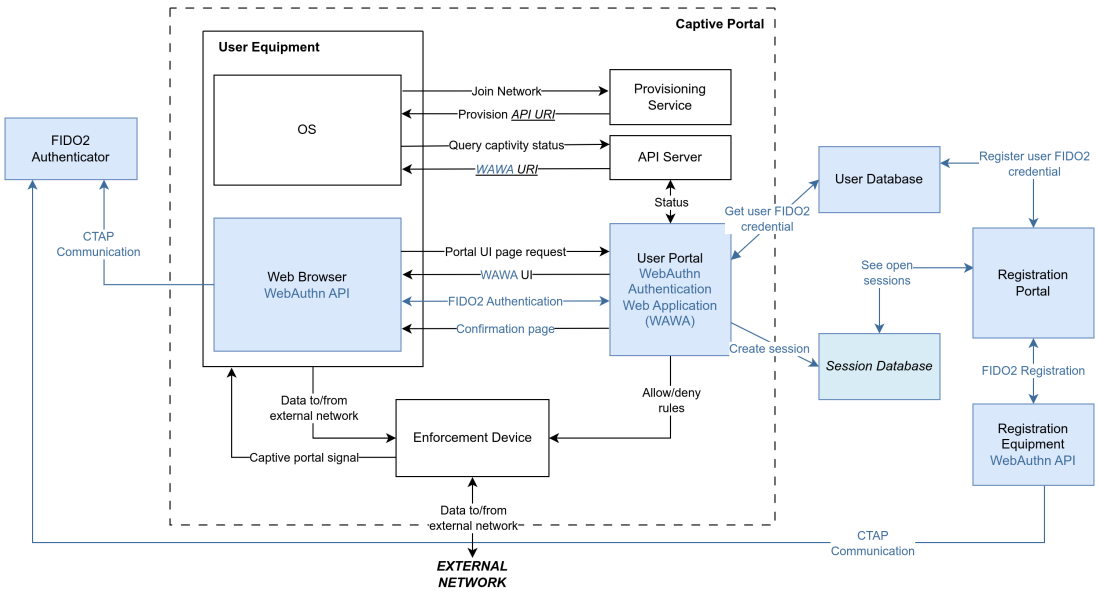


Fig. 1. Architecture of the FIDO2 Captive-portal Authentication (FIDO2CAP) Protocol.

authentication process has been embeded in the captive portal operation.

After the captive portal signaling the User Equipment, the WAWA URI is provided to the client, leading the user to the WAWA interface. The user submits their username, allowing the WAWA to retrieve associated FIDO2 Authenticators from the User Database. A WebAuthn challenge is generated and sent to the browser, which interacts with the authenticator to complete the authentication. Successful verification prompts the Enforcement Device to grant network access.

III. Results and conclusion

The first result is the FIDO2CAP protocol, a novel network authentication solution based on FIDO2 integrated with captive portals. The protocol architecture adapts RFC 8952 [4], incorporating elements like the WebAuthn Authentication Web Application (WAWA) and FIDO2 authenticators. Secondly, a prototype was developed and deployed in a mock hotel Wi-Fi network environment, offering a functional implementation with compatibility across diverse devices. Finally, compatibility and usability tests were conducted with 15 real users, demonstrating that over 69% successfully completed authentication, with most finding the system easy to use, despite encountering some errors during the process.

The usability study revealed several types of errors, including user, OS, and connectivity issues, but these did not prevent completion. In fact, users performed better on subsequent attempts, showing an improvement in efficiency. Despite these errors, 66.67% of participants expressed willingness to use security keys for Wi-Fi authentication, indicating general acceptance. However, issues such as the reliance on operating system-specific mini-browsers and concerns about physical token loss highlighted areas for improvement in the user experience and broader adoption.

FIDO2CAP offers a robust and secure alternative to traditional password-based systems, directly integrating FIDO2 authentication into network access control. This solution contrasts with other approaches, like temporary password provisioning, by eliminating the need for additional applications or credentials. Although the protocol shows potential, further work is needed to enhance software compatibility, address usability concerns, and promote adoption across diverse platforms. The success of FIDO2CAP in this pilot study lays the foundation for future research and development in secure, passwordless authentication systems.

Acknowledgements

This work was supported by the grant ED431C 2022/46 – Competitive Reference Groups GRC – funded by: EU and “Xunta de Galicia” (Spain). This work was also supported by CITIC, funded by “Xunta de Galicia” through the collaboration agreement between the “Consellería de Cultura, Educación, Formación Profesional e Universidades” and the Galician universities to strengthen the research centres of the “Sistema Universitario de Galicia” (CIGUS). Also, the work is founded by the “Formación de Profesorado Universitario” (FPU) grant from the Spanish Ministry of Universities to Martiño Rivera Dourado (Grant FPU21/04519).

References

[1] M. Rivera-Dourado, M. Gestal, A. Pazos, and J. A. Vázquez-Naya, “A novel protocol using captive portals for FIDO2 network authentication,” Appl. Sci., vol. 14, no. 9, Art. no. 3610, 2024. doi:10.3390/app14093610.

[2] Cisco Systems, “The 2022 Duo Trusted Access Report,” 2022. Available online: https://duo.com/assets/ebooks/the-2022-duo-trusted-access-report.pdf

[3] Web Authentication: An API for Accessing Public Key Credentials Level 3. Available online: https://www.w3.org/TR/webauthn-3/.

[4] Larose, K.; Dolson, D.; Liu, H. RFC 8952: Captive Portal Architecture. RFC Editor. 2020. doi: 10.17487/RFC8952.

A review of: Credenciales Colaborativas para el Internet de las Cosas

Santiago de Diego
TECNALIA, Basque Research and Technology Alliance (BRTA),
Derio, 48160, España
santiago.dediego@tecnalia.com

Cristina Regueiro
TECNALIA, Basque Research and Technology Alliance (BRTA),
Derio, 48160, España
cristina.regueiro@tecnalia.com

Gabriel Maciá-Fernández
Departamento de Teoría de la Señal- CITIC, Telemática y Comunicaciones, Universidad de Granada,
Granada, 18071, España
gmacia@ugr.es

Resumen- **El concepto de Credencial Colaborativa (CC) ha surgido como una solución innovadora para abordar las limitaciones de los protocolos estándar de Identidad Auto-Soberana (SSI) en la gestión de identidades grupales dentro del ecosistema del Internet of Things (IoT). Las CCs permiten implementar casos de uso que requieren colaboración y que no pueden resolverse con credenciales verificables (VCs) estándar de SSI, aumentan la privacidad de los participantes del grupo y permiten el desarrollo de un marco de software que cualquier verificador podría usar para generar una aplicación genérica. Este artículo introduce las CCs como un procedimiento formalizado de gestión de identidad que permite la colaboración entre múltiples participantes para cumplir con los requisitos del verificador y resume los principales hallazgos del artículo [2].**

***Index Terms*- SSI, Identidad Auto-Soberana, IoT**

Tipo de contribución: *Investigación ya publicada*

I. Introducción

El Internet de las Cosas (IoT) está experimentando un rápido crecimiento, que trae consigo nuevos desafíos en la gestión de identidad, particularmente debido a la gran cantidad de dispositivos y su naturaleza a menudo limitada. Si bien la Identidad Auto-Soberana (SSI) ha logrado avances significativos en la descentralización de la gestión de identidad, los estándares actuales, como el de W3C [1] no abordan adecuadamente escenarios donde múltiples entidades y o dispositivos necesitan colaborar para lograr un objetivo común. El presente artículo es un resumen del artículo publicado [2], el cual presenta el concepto de Credenciales Colaborativas y su aplicación al IoT. Por ende, tiene como objetivo principal presentar las CCs, su ciclo de vida, y sus conclusiones más relevantes.

II. Credenciales Colaborativas

Las CC se definen como un grupo de credenciales verificables (VCs) generadas a través de la colaboración entre múltiples participantes. Los beneficios clave de las CC incluyen: i) habilitar casos de uso que requieren colaboración y que no pueden resolverse con VCs estándar, ii) aumentar la privacidad de los participantes del grupo, al permitir exponer únicamente la información que se desea revelar y delegar la creación de la CC en la figura del coordinador del grupo, y iii) facilitar el desarrollo de un marco de software genérico para cualquier verificador. Las CC son particularmente útiles en escenarios de IoT donde los dispositivos pueden colaborar para demostrar propiedades que requieren la contribución de varios participantes (por ejemplo, consumos agregados de un grupo de sensores).

Aunque existen conceptos relacionados en el estado del arte, como las Credenciales de Grupo [3], las CCs se diferencian por su enfoque dinámico y su funcionamiento en tiempo real. Mientras que las Credenciales de Grupo identifican a un conjunto predefinido de dispositivos o usuarios, las CCs se generan en base a la interacción activa y coordinada entre los participantes dentro de un contexto específico y dinámico. Las CC hacen uso de la delegación de credenciales, la cual es especificada por la W3C [1].

A. *Modelo formal*

Formalmente, una CC puede estar formada por: i) VCs firmadas y emitidas por los propios participantes del grupo que certifican a otros participantes, ii) VCs firmadas por varios participantes (con un proceso de firma múltiple), iii) VCs firmadas por emisores externos con *claims* referentes a los participantes del grupo, iv) VCs delegadas firmadas por el *subject* original y enviadas a otro participante, v) CCs, ya que que una CC podría a su vez ser parte de otra CC en una colaboración entre grupos.

B. *Casos de uso de las Credenciales Colaborativas*

En [2] se discuten varias aplicaciones potenciales para las CC, incluyendo: i) escenarios de IoT como Servicio [4], como centros de emergencia que requieren datos de múltiples sensores, ii) control de acceso para dispositivos múltiples basado en la verificación del equipo, iii) controles de seguridad en las industrias de aviación y automotriz, iv) prevención de ataques de suplantación de identidad en dispositivos IoT. Estos casos de uso destacan la versatilidad y el impacto potencial de las CC en diversas industrias.

III. Ciclo de vida de las CCs

A continuación, se presentan los roles que participan en el ciclo de las CC. Dichos roles son:

Participantes: Titulares de credenciales que colaboran para crear una CC.

Coordinador: Participante elegido para liderar el proceso de creación de la CC.

Verificador: Entidad que solicita y verifica la CC.

Emisor: Entidad que emite las credenciales individuales a los participantes.

El ciclo de vida de una CC incluye las siguientes fases:

Inicialización:

Primero de todo, el verificador define y publica los requisitos para la CC, y posteriormente, los participantes potenciales descubren estos requisitos. Esto se realiza mediante lo que se conoce como Especificación de Condiciones (CS), la cual permite saber al grupo cuales son los requisitos que necesita el verificador, a fin de poder satisfacerlos a nivel grupal. El formato completo, así como los campos que debe contener esta CS pueden consultarse en [2].

Descubrimiento:

El protocolo de descubrimiento permite al coordinador descubrir e identificar a los posibles miembros del grupo.

Incorporación:

Una vez que el descubrimiento ha finalizado, se requiere implementar otro protocolo, denominado Incorporación. Este protocolo se utiliza para establecer una comunicación uno a uno entre los miembros del grupo. En general, la decisión de unirse o no al grupo dependerá enteramente del caso de uso, ya que a veces los miembros del grupo podrían verse obligados a unirse, gestionados por el coordinador, de modo que los miembros del grupo podrían actuar como esclavos, siguiendo las órdenes del maestro.

Selección del coordinador:

Los participantes eligen un coordinador mediante un algoritmo de consenso. Posteriormente, el coordinador asume la responsabilidad de recopilar y agregar la información relativa al grupo.

Recopilación de credenciales:

Cada participante genera presentaciones verificables (VPs) basadas en sus VCs individuales y dichas VPs se envían al coordinador de forma segura.

Generación de la CC:

El coordinador agrega las VPs recibidas, se crea la estructura de la CC con todos sus componentes y el coordinador finalmente presenta la CC al verificador. El verificador valida la CC comprobando las firmas y el cumplimiento de los requisitos.

IV. Conclusiones de la evaluación del rendimiento

En el artículo original ([2]) se proporciona un ejemplo de implementación utilizando Hyperledger Aries e Indy, demostrando la viabilidad del modelo de CC y de su ciclo de vida propuesto. Además de su implementación, se han realizado una serie pruebas con la finalidad que analizar el impacto que tienen las CCs en el sistema final. Se ha medido también cómo escala el sistema propuesto cuando el número de participantes aumenta, y cual es el impacto energético que tendría realizar todo el flujo propuesto en la Fig. 1.

Las conclusiones son las siguientes: i) los procesos de descubrimiento e incorporación tienen el impacto más significativo en el rendimiento, ii) El protocolo escala bien (incremento lineal del tiempo) con un número creciente de miembros del grupo, iii) la adición de un coordinador no aumenta significativamente el tiempo del protocolo mientras proporciona numerosas ventajas, como se ha visto, iv) en cuanto al gasto energético, en una batería normal de 5200 mAh, dicha batería instalada en uno de los participantes ejecutando constantemente el protocolo duraría alrededor de 11,36 días sin carga, lo cual es aceptable en un entorno como el considerado.

Finalmente, se ha realizado un análisis de seguridad del sistema propuesto. Entre los ataques analizados se incluyen

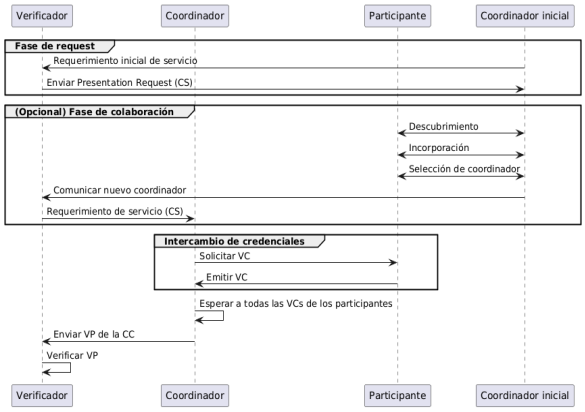


Fig. 1. Flujo completo de las CCs (basado en el flujo propuesto en [2]).

retrasos en el envío de VCs, ataques de denegación de servicio (DoS), manipulación de la selección del coordinador, suplantación de identidad y ataques Sybil. Se proponen soluciones como el uso de mecanismos de tiempo de espera, algoritmos de consenso tolerantes a fallos bizantinos, *Trusted Execution Environments* (TEEs) o *Trusted Platform Modules* (TPMs). Finalmente, se menciona la importancia de tener una lógica adicional para situaciones de emergencia, como la capacidad de anular manualmente el sistema en casos críticos.

V. Conclusiones

El artículo concluye que las CC ofrecen una solución novedosa para gestionar identidades grupales en escenarios de IoT, abordando limitaciones clave de los protocolos SSI actuales, permitiendo la colaboración en tiempo real y mejorando la privacidad. Las líneas de investigación futuras incluyen: i) desarrollar mecanismos de consenso óptimos para los miembros del grupo, ii) explorar protocolos de votación eficientes para el coordinador, iii) optimizar los procesos de recreación de grupos, iv) investigar patrones de diseño para distintas aplicaciones basadas en CC.

Agradecimientos

Este trabajo se ha desarrollado dentro del “Plan de Recuperación, Transformación y Resiliencia”, convenio C025/24 INCIBE-Universidad de Granada, financiado por la Unión Europea (fondos Next Generation) y del proyecto NetSEA-GPT (C-ING-300-UGR23), cofinanciado por la Consejería de Universidad, Investigación e Innovación y la Unión Europea (Programa FEDER Andalucía 2021-2027).

Referencias

[1] World Wide Web Consortium: “Verifiable Credentials Data Model 1.0: Expressing Verifiable In-formation on the Web”, en *https://www. w3. org/TR/vc-data-model/?#core-data-model*.

[2] de Diego, S., Regueiro, C., & Maciá-Fernández, G: “Collaborative credentials for the Internet of Things”, en *Computer Networks*, vol. 251 y pp. 110629, 2024.

[3] Fotiou, N., Siris, V. A., Xylomenos, G., & Polyzos, G. C: “IoT Group Membership Management Using Decentralized Identifiers and Verifiable Credentials”, en *Future Internet*, vol. 14 (6) y pp. 173, 2022.

[4] De Diego, S., Regueiro, C., & Maciá-Fernández, G: “Enabling identity for the IoT-as-a-service business model”, en *IEEE Access*, vol. 9 y pp. 159965-159975, 2021.

El delito de estafa informática y el *phishing* tras la reforma de la LO 14/2022. La ciberseguridad como reto

Dra. María Teresa Montero Ruiz*
Departamento de Derecho Público y de la Empresa
Facultad de Ciencias Jurídicas y Sociales de Toledo
Universidad de Castilla-La Mancha
Teresa.montero@uclm.es

Dra. María Pilar Marco Francia
Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Derecho Procesal y Derecho Penal
Facultad de Derecho Universidad Complutense de Madrid
Pilar.marco@ucm.es

Resumen- La ciberdelincuencia, especialmente el *phishing*, ha crecido en la sociedad digital, lo que llevó a modificar la Ley Orgánica 14/2022 del Código Penal español, específicamente el artículo 249. Esta reforma busca regular mejor la estafa informática y alinearse con normativas europeas, protegiendo a los consumidores y la integridad del mercado digital. El *phishing*, que obtiene información confidencial mediante técnicas fraudulentas, ha evolucionado en diversas modalidades como *smishing* y *vishing*. El nuevo artículo 249 tipifica acciones que resultan en la transferencia no consentida de activos, con penas de 6 meses a 3 años de prisión. En la banca online, la legislación establece responsabilidades para las entidades financieras ante fraudes, aunque inicialmente fueron eximidas de culpa. Sin embargo, recientes sentencias han comenzado a responsabilizarlas por falta de medidas de seguridad adecuadas. La ciberseguridad es clave para mitigar el *phishing*, requiriendo un enfoque integral con controles técnicos y tecnologías avanzadas como la inteligencia artificial. La reforma de 2022 es un avance en la lucha contra la ciberdelincuencia, pero es esencial que el marco normativo se adapte a las nuevas realidades del delito informático. La colaboración internacional y medidas preventivas son vitales para proteger a los consumidores y el mercado digital.

Index Terms- Ciberdelincuencia, *phishing*, estafa informática, muleros

Tipo de contribución: Investigación original (límite 8 páginas).

I.- Introducción

La ciberdelincuencia se ha convertido en una de las preocupaciones más serias en nuestra sociedad digital actual. A medida que la tecnología avanza, las tácticas empleadas por los delincuentes han cambiado, lo que ha llevado a un aumento notable de prácticas en los últimos años. En este sentido, la modificación de la Ley Orgánica 14/2022, del 22 de diciembre, ha influido de manera considerable en la forma en que se regulan estos delitos en el Código Penal (CP), en particular en el artículo 249.

Es fundamental abordar el problema de la estafa informática, especialmente el *phishing*, dado que cada vez más personas son susceptibles a las artimañas que utilizan los estafadores. La reforma legal busca no solo definir con mayor claridad estas conductas, sino también alinearse con las normativas europeas, como la Directiva (UE) 2019/713, que tiene como objetivo unificar las legislaciones de los países miembros en cuestiones de fraude y falsificación de métodos de pago no en efectivo.

Nos proponemos analizar cómo se define el delito de *phishing* en el nuevo artículo 249 del CP, así como las consecuencias de esta reforma, las distintas modalidades delictivas que pueden surgir y la responsabilidad penal de todos los involucrados, incluidos los conocidos como "muleros". Además, es crucial revisar las sentencias más relevantes que han abordado estos nuevos artículos y considerar las implicaciones prácticas para la lucha contra estos delitos en un entorno cada vez más digital.

II.- La estafa informática en el CP

La estafa ha sido, en términos generales, un delito reconocido en el CP desde su creación y la estafa informática se encuadraba en el delito de estafa general, que exige la presencia de un engaño. Sin embargo, con la llegada de la era digital, han surgido nuevas modalidades que no se ajustan a las definiciones tradicionales contenidas en el artículo 248 del CP, pues, en estos casos, el desplazamiento de bienes sin consentimiento no se produce a través del engaño a una persona física, sino mediante tácticas fraudulentas que utilizan manipulaciones informáticas o técnicas diseñadas para defraudar a terceros, facilitadas por las herramientas tecnológicas actuales [**1**].

ISBN:978-84-10169-61-6

En todo caso, el artículo 248 del CP describía la estafa y sus diversas modalidades, incluyendo transferencias no autorizadas y el uso fraudulento de tarjetas, mientras que el artículo 249 establecía las penas, clasificando como delito leve aquellas estafas que no superaran los 400 euros. Con la entrada en vigor de la Ley Orgánica 14/2022, de 22 de diciembre, que entró en vigor el 12 de enero de 2023, se modificaron estos artículos. En la actualidad, el artículo 248 define la estafa básica y continúa considerando leves los fraudes por debajo de los 400 euros, aunque no abarca casos especiales. Por su parte, el artículo 249 se centra en las estafas informáticas, incorporando nuevos supuestos y manteniendo algunos del artículo 248, pero sin especificar la cuantía. Esto plantea el debate sobre la posible desproporción de imponer penas más severas por estafas de menor cuantía, si se considera únicamente el medio utilizado, dado que la cuantía es un criterio habitual en los delitos contra el patrimonio.

Así pues, esta reforma introdujo cambios significativos en el artículo 249 del CP, que ahora aborda de manera específica el delito de estafa informática, buscando no solo tipificar el delito de *phishing*, sino también incluir actos preparatorios que faciliten la comisión de delitos relacionados con medios de pago distintos del efectivo.

Todo ello surgió de la necesidad de transponer la Directiva 2019/713 es un documento clave en la regulación del fraude y la falsificación de medios de pago. Establece la necesidad de un enfoque común en el derecho penal para castigar conductas que, aunque pueden variar en su forma, comparten el mismo objetivo: la protección de los consumidores y la integridad del mercado digital. La Directiva subraya que el fraude y la falsificación de medios de pago son obstáculos para el mercado único digital y afectan la confianza de los consumidores.

III.- El *phishing*

El *phishing*, también conocido como "pescado de datos", es una de las técnicas más comunes empleadas por delincuentes informáticos en la actualidad. El término proviene de las palabras en inglés "*password harvesting*" (cosecha de contraseñas) y "*fishing*" (pesca), refiriéndose a la acción de obtener información confidencial de los usuarios. Desde la década de 1990, esta técnica ha evolucionado, comenzando con el envío masivo de correos electrónicos fraudulentos a individuos y empresas. Existen variantes como el *smishing* y el *vishing*, que utilizan mensajes de texto y llamadas telefónicas, respectivamente. Este fenómeno se desarrolla en diversos territorios y puede involucrar múltiples países, lo que da percepción de la complicada labor que enfrentan las fuerzas y cuerpos de seguridad al abordar delitos que pueden ocurrir en momentos y lugares específicos [**2**]. En todo caso, la doctrina del TS establece que la competencia judicial para juzgar los casos de *phishing* se determina por la ubicación de los elementos del delito, aplicando la teoría de la ubicuidad, que indica que el delito se comete en todos los lugares donde se desarrollan sus elementos, permitiendo que el juzgado que primero inicie las actuaciones tenga competencia.

Los ataques de *phishing* se centran en obtener credenciales de acceso a diversas plataformas, como redes sociales y correos electrónicos, aprovechándose de la confianza que generan las entidades suplantadas. Con el tiempo, los métodos han cambiado, y ahora también incluyen publicaciones en redes sociales que incitan a las víctimas a proporcionar información personal a cambio de beneficios.

Los delincuentes suelen enviar correos que parecen proceder de instituciones financieras legítimas, utilizando imágenes y páginas web que imitan a las de estas entidades para dificultar su identificación como fraudes. A través de atractivas ofertas o solicitudes de contraseñas, buscan acceder a las cuentas de los usuarios. Este enfoque se asemeja a "pesca con red", ya que se dirige a un amplio público, esperando que al menos algunos caigan en la trampa. Una vez que los delincuentes obtienen las credenciales de acceso, pueden transferir fondos o vender la información en el mercado negro, complicando la persecución del delito debido a las dificultades técnicas y a las complejidades jurisdiccionales en el Derecho penal. Además, las instituciones financieras a menudo, para proteger su reputación, asumen la responsabilidad de compensar a las víctimas del fraude a menos que se demuestre negligencia por parte del cliente.

Lo cierto es que no hay un único *phishing*, sino múltiples [**3**], como el *phishing* tradicional, que es la modalidad más simple desde el punto de vista técnico, pues consiste en crear una copia de un sitio web familiar para la víctima, reemplazando la dirección original por una falsa. Cuando el usuario ingresa sus datos en esta página emulada, la información es capturada y puede ser almacenada en un archivo de texto o enviada a una dirección de correo electrónico del atacante. Este tipo de *phishing* se limita principalmente a la suplantación de un sitio web. El *Spear Phishing*, que está dirigido a un grupo específico de personas, utilizando técnicas de ingeniería social y un estudio detallado de las víctimas para personalizar los ataques [**4**]. En este caso, se envían correos electrónicos que incluyen información específica que hace que el mensaje parezca legítimo. Los atacantes buscan el eslabón más débil dentro de una organización, dirigiéndose a personas que no tienen conocimientos técnicos para facilitar el engaño. El *Smishing* (*Phishing* por SMS)**,** que utiliza mensajes de texto en teléfonos móviles para engañar a los usuarios. Los mensajes suelen falsificar enlaces a sitios web y pueden ofrecer premios o beneficios, solicitando información personal a cambio. Además, se ha expandido a aplicaciones de mensajería como WhatsApp, donde los mensajes pueden enviarse sin costo adicional, siempre que haya conexión a Internet. El *Vishing* (*Phishing* por voz)**,** más reciente, que se basa en el uso de la telefonía IP. Los atacantes imitan el nombre de marcas reconocidas, como bancos, y solicitan a las víctimas que llamen a un número específico. Al hacerlo, la persona es atendida por un sistema automatizado que busca recopilar información personal. Estos falsos centros telefónicos suelen estar conectados entre sí para aumentar la credibilidad del engaño y, por ende, su efectividad. El *Tabnabbing*, que se fundamenta en la falta de atención de los usuarios que mantienen múltiples pestañas abiertas simultáneamente durante su navegación en Internet. Los ciberdelincuentes, al notar que una pestaña ha estado inactiva durante un tiempo, aprovechan esta situación para modificar su contenido, haciéndola parecer como si fuera la original, con el objetivo de inducir a los usuarios a ingresar nuevamente en ella y

consumándose así la estafa. Y el *Pharming* que produce la estafa cuando se instala un código malicioso en el ordenador para redirigir al usuario a sitios web falsos.

IV.- Tipificación del *phishing* en el CP

El artículo 249.1 a) del CP tipifica el delito de estafa informática en su modalidad de *phishing*, castigando con la pena de 6 meses a 3 años de prisión a "*los que, con ánimo de lucro, obstaculizando o interfiendo indebidamente en el funcionamiento de un sistema de información o introduciendo, alterando, borrando, transmitiendo o suprimiendo indebidamente datos informáticos o valiéndose de cualquier otra manipulación informática o artificio semejante, consigan una transferencia no consentida de cualquier activo patrimonial en perjuicio de otro"*. Así, los elementos constitutivos del delito incluyen el ánimo de lucro, ya que el autor debe actuar con la intención de obtener un beneficio económico; la manipulación informática, pues se requiere que la conducta implique la alteración, eliminación o transmisión indebida de datos; y la transferencia no consentida de activos, dado que el resultado de la acción debe ser la obtención de un activo patrimonial en perjuicio de otro.

La jurisprudencia ha jugado un papel clave en la interpretación de la nueva redacción del artículo 249 del CP, estableciendo el Tribunal Supremo (TS) que, en el caso de la estafa informática, no es necesario probar el engaño personal, ya que la manipulación informática actúa de manera automática en perjuicio de la víctima.

V.- Actores del *Phishing*

El *phishing* involucra al "*phisher*", persona que realiza el fraude, el "mulero" que facilita el movimiento del dinero defraudado y la víctima.

Los *phishers*, es decir, aquellos que perpetran el ataque y se presentan como individuos o instituciones en las que la víctima confía, como un compañero de trabajo, una institución financiera o cualquier otra organización familiar para la persona afectada. Mediante el envío de correos electrónicos, solicitan que la víctima acceda a su cuenta o proporcione información confidencial en un sitio web fraudulentamente creado que imita a uno auténtico, todo ello, con el objetivo de robar credenciales o información de tarjetas de crédito.

La figura del "mulero" es clave en la dinámica delictiva, ya que permite a los estafadores ocultar su identidad y el origen del dinero, pues es la persona que actúa como intermediario, recibiendo fondos obtenidos de manera fraudulenta y transfiriéndolos a los estafadores a cambio de una comisión.

La doctrina ha identificado varias formas de calificar penalmente la conducta del "mulero", así se le puede considerar autor del delito de estafa informática por su papel activo en la transferencia de fondos. En algunos casos, su conducta puede encuadrarse en el delito de blanqueo, especialmente si actúa con conocimiento del origen ilícito de los fondos. También se ha argumentado que el "mulero" puede incurrir en delito de receptación si ayuda a los responsables a aprovecharse de los efectos del delito.

Los tribunales han abordado la responsabilidad del "mulero" en múltiples sentencias, estableciendo que su participación en la estafa informática puede considerarse cooperación necesaria, dado que su intervención es esencial para la consumación del delito. En todo caso, si hablamos de la participación intencionada del "mulero", es fundamental que conozca perfectamente qué tipo de ayuda está ofreciendo, estableciéndose su responsabilidad sólo si se puede demostrar que actuó con pleno conocimiento de que su conducta estaba contribuyendo a un delito específico. Sin embargo, en muchas ocasiones, bajo la teoría de la "ignorancia deliberada" [**5**], se estén penalizando situaciones que podrían considerarse como dolo eventual, pero también otras que se presentan como cuestionables en cuanto a su carácter doloso. Esto pone de manifiesto una representación insuficiente de todos los elementos que conforman el delito en cuestión. Si esta falta de representación es total, nunca podrá sostenerse una imputación subjetiva bajo el concepto de dolo.

Respecto a la posición de la víctima, si bien a veces proporcionan información que facilita el fraude, el TS ha rechazado la idea de que tengan una obligación de autoprotección ante engaños complejos.

VI.- El *phishing* y la banca *online*

El artículo 41.1 del Real Decreto 19/2018, de 23 de noviembre, sobre servicios de pago, los usuarios tienen la obligación principal de adoptar medidas razonables para proteger sus credenciales. Sin embargo, este Real Decreto también establece un sistema de responsabilidad cuasi-objetiva para la entidad proveedora del servicio de pago, donde se invierte la carga probatoria, presumiendo la falta de autorización si el titular lo niega. Esta responsabilidad civil solo se extingue si el cliente actúa fraudulentamente o con negligencia grave en la protección de sus datos. Además, el cliente debe informar a la entidad sobre cualquier pago no autorizado en cuanto tenga conocimiento, siempre que la entidad disponga de un sistema de comunicación adecuado, gratuito y accesible en todo momento.

Efectivamente, el campo de la banca online, la Sentencia del TS del 2 de diciembre de 2014 establece un régimen de cuasi-responsabilidad civil objetiva para los bancos, que se fundamenta en la falta de diligencia por parte de las entidades bancarias que no implementaron las medidas adecuadas para proteger a sus clientes en el servicio. Como resultado, se eliminó la obligación de la víctima de protegerse a sí misma, excepto en casos de fraudes que sean evidentes para cualquier persona. Por lo tanto, es responsabilidad del banco adoptar las medidas necesarias para prevenir este tipo de delitos. No obstante, las primeras resoluciones sobre estos fraudes generalmente han fallado a favor de los bancos, eximiéndolos de responsabilidad frente a los clientes que entregaron sus contraseñas de manera voluntaria.

Sin embargo, la Audiencia Provincial de Badajoz en sentencia de 14 de febrero de 2025, obliga a una entidad bancaria a indemnizar a una clienta víctima de un fraude de *phishing* al no haber implementado medidas adecuadas para prevenir este tipo de engaños. El incidente se produjo cuando

la clienta, al acceder a su aplicación bancaria, recibió un mensaje que le indicaba la necesidad de activar un nuevo sistema de seguridad. Así, al seguir las instrucciones, ingresó sus credenciales en una página falsa, lo que resultó en la sustracción de sus fondos. Si bien el Juzgado de Primera Instancia argumentó que la clienta era responsable del fraude, esta afirmación fue rechazada por la Audiencia que subrayó que los consumidores, especialmente aquellos menos familiarizados con la tecnología, tienen dificultades para detectar fraudes cada vez más sofisticados, de forma que los bancos, como custodios del dinero de sus clientes, deben asumir la responsabilidad en tales incidentes. Además, enfatiza la necesidad de mejorar las comunicaciones seguras, ofreciendo formación adecuada a los usuarios y realizando inversiones en seguridad para prevenir la clonación de páginas web. Destaca que ser víctima de un fraude no implica negligencia por parte del cliente, y que la responsabilidad recae en el banco a menos que se demuestre una negligencia extrema por parte del cliente.

En definitiva, el banco deberá asumir su responsabilidad cuando no establece medidas de seguridad apropiadas, cuando desestima indicios de fraude como transacciones inusuales, y cuando, al recibir una queja del cliente, no toma acciones rápidas para frenar las transferencias o lo hace con demora. También cuando no realice comprobaciones adecuadas con los usuarios en relación a operaciones que generen dudas.[**6**]

VII.- Ciberseguridad contra el *Phishing*

En la actualidad, la protección cibernética se ha vuelto esencial tanto para empresas como para individuos. La salvaguarda de la información es uno de los aspectos más críticos, ya que cualquier vulnerabilidad puede tener repercusiones severas. Entre las amenazas que pueden poner en riesgo la integridad de los datos, se encuentran los errores humanos, problemas en la transferencia de datos, *software* malicioso [**7**] y fallos en el *hardware*.

Uno de los problemas más sutiles es el error humano, que puede presentarse de diversas maneras, como la duplicación de información, la eliminación accidental de datos o la introducción errónea de información. Un simple descuido puede conducir a decisiones equivocadas basadas en datos incorrectos, lo que podría afectar notablemente tanto la operativa de una empresa como la vida personal de un individuo.

Los errores en la transferencia de información también representan un reto considerable. Es fundamental que los datos se mantengan consistentes al moverse de una base de datos a otra. Sin embargo, si ocurren fallos durante este proceso, puede haber discrepancias entre las bases de datos de origen y destino, lo que ocasiona confusión y dificultades en la gestión de la información.

Asimismo, el aumento en la circulación de virus y errores de programación, como *spyware*, *malware* y troyanos, agrava la situación. Estos programas dañinos pueden infiltrarse en los sistemas, amenazando no solo la integridad de los datos, sino también la confidencialidad y disponibilidad de la información. Esta creciente amenaza resalta la importancia de mantener actualizados los sistemas de seguridad y de educar a los usuarios sobre los peligros que enfrentan.

Por último, el *hardware* comprometido representa otra fuente de riesgo. Fallas inesperadas en computadoras o servidores pueden resultar en la pérdida de datos cruciales y causar inconvenientes operativos. Es fundamental realizar un mantenimiento regular del equipo para reducir estos riesgos.

Para enfrentar la amenaza del *phishing*, que pone en jaque la integridad de los datos y la confianza en las comunicaciones digitales, es vital establecer una serie de controles y medidas técnicas que protejan a las organizaciones de estos ataques maliciosos.

La estrategia de ciberseguridad contra el *phishing* debe ser integral y combinar diversas medidas para salvaguardar la información. Algunas de las tácticas incluyen: el uso de algoritmos de *hashing*, que generan una huella digital única de los datos, que aunque no permiten detectar modificaciones por sí mismos; la gestión efectiva de configuraciones, que implica monitorear constantemente las configuraciones y aplicar políticas que aseguren que cualquier cambio sea debidamente aprobado y documentado; y un control exhaustivo de la gestión de cambios, que garantiza que cada modificación sea evaluada y revisada para evitar errores y vulnerabilidades.

Además, es esencial establecer controles de acceso físico que restrinjan la entrada a las instalaciones a personal autorizado (aunque no contribuya significativamente a paliar los ataques de *phishing*), complementados por controles de acceso lógico que limiten el acceso a información y aplicaciones sensibles. Las políticas de autenticación robustas son clave para asegurar que solo los usuarios autorizados puedan acceder a recursos críticos. La firma digital también juega un papel importante, ya que permite verificar la autenticidad e integridad de la información mediante criptografía, asegurando que solo el remitente autorizado pueda enviar datos y que estos no hayan sido alterados en el proceso.

Respecto a la Inteligencia Artificial (IA), también puede desempeñar un papel crucial en la detección y prevención de ataques de ingeniería social [**8**], pues emplea técnicas avanzadas de identificación de *phishing* que analizan el contenido de los correos electrónicos, empleando algoritmos de aprendizaje automático para reconocer patrones que indican la posible maliciosidad de un mensaje. La IA es capaz de detectar perfiles fraudulentos en redes sociales y otras plataformas digitales, lo que contribuye a prevenir ataques de ingeniería social antes de que ocurran.

No obstante, también existe el riesgo de que la IA sea utilizada por los ciberatacantes con fines malintencionados, como la creación de perfiles falsos en redes sociales y aplicaciones de citas, con el objetivo de contactar a las víctimas y obtener información personal. Además, los ataques de *phishing* están evolucionando y resultan cada vez más sofisticados al emplear técnicas de generación de texto mediante IA, lo que les permite crear correos electrónicos que aparentan ser más auténticos y persuasivos para las potenciales víctimas.

En todo caso, implementar protocolos que incluyan comprobaciones de redundancia durante la transmisión de datos es crucial para garantizar la integridad de la información, utilizando técnicas como el envío de paquetes duplicados y la verificación de que todos los datos se hayan recibido correctamente, lo que reduce el riesgo de pérdida o corrupción de información.

VIII.- Conclusiones

La reforma de la LO 14/2022 ha permitido una mejor adecuación del ordenamiento jurídico español a la realidad de los delitos informáticos, especialmente en lo que respecta al *phishing* y la estafa informática. La nueva regulación no solo tipifica de manera más precisa estas conductas, sino que también establece un marco legal que permite la persecución efectiva de los delitos informáticos y sus actos preparatorios.

Los "muleros" representan un desafío adicional para el sistema penal, dado que su papel en la cadena delictiva puede variar en función de su conocimiento y participación en el delito. La jurisprudencia ha sido clara en la necesidad de considerar su intervención como esencial para la consumación de la estafa, permitiendo así la imposición de penas adecuadas.

En un mundo cada vez más digitalizado, es fundamental que el marco legal continúe evolucionando para abordar las nuevas formas de criminalidad y proteger a los consumidores y a la integridad del mercado. La colaboración entre Estados miembros y la implementación de medidas preventivas y reactivas son esenciales para combatir eficazmente la estafa informática y sus modalidades, como el *phishing*.

AGRADECIMIENTOS

El presente trabajo se ha realizado con los fondos del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation), a través de la Cátedra "Ciberseguridad para la Innovación y la Protección Digital" INCIBE-UCM.

Además, este trabajo ha contado con el apoyo de los siguientes Proyectos de Investigación:

Proyectos de la Comisión Europea: HEROES (acuerdo de subvención nº. 101021801) y ALUNA (https://aluna-isf.eu/), en virtud del acuerdo de subvención nº 101084929. El contenido de este artículo no refleja la opinión oficial de la Unión Europea ni de la Comisión Europea. Los autores son los únicos responsables de la información y las opiniones expresadas en él.

Proyecto «Violencia extrema sobre la mujer: respuesta penal y alternativas para la prevención» (VEM, PID2022-140170NB-I00). 2022-2026. CO-IP's Carolina Bolea Bardon y Juan Carlos Hortal Ibarra. Entidad financiadora: Ministerio de Ciencia e Innovación del Gobierno de España.

Proyecto Criminalidad organizada transnacional y empresas multinacionales ante las vulneraciones a los derechos humanos. Nº: PID2020-117403RB-100. IP Laura Zúñiga Rodríguez. Programa financiador: Programa Estatal de I+D+I orientada a los Retos de la Sociedad, del Plan Estatal de Investigación Científica y Técnica y de Innovación de 2017-2020. Entidad financiadora: Ministerio de Ciencia e Innovación del Gobierno de España.

"Cuestiones controvertidas en relación con el ejercicio de la capacidad jurídica por los y las menores de edad" (Ref. CIGE/2023/71), financiado por la Conselleria de Educación, Universidades y Empleo de la Generalitat de Valencia. I.P. P. Chaparro Matamoros."

REFERENCIAS

[**1**] Núñez Castaño, E. "Delitos patrimoniales de enriquecimiento mediante defraudación (I) estafa", En *Nociones fundamentales de derecho penal*. Gómez Rivero (dir.) Tecnos, 2019.

[**2**] Rodríguez González, Víctor, Payá Santos, Claudio Augusto, y Peña Herrera, Bernardo. "Estudio criminológico del ciberdelincuente y sus víctimas". *Cuadernos de RES PUBLICA en derecho y criminología, nº 2,* pp. 95–107, 2023.

[**3**] Miró Llinares, F. "La respuesta penal al ciberfraude: Especial atención a la responsabilidad de los muleros del phishing", *Revista electrónica de ciencia penal y criminología, n.º 15,* 2013.

[**4**] Rodríguez Almirón, F. "El delito de estafa informática. ¿Es posible determinar la responsabilidad civil de la entidad financiera en base al artículo 120.3 del Código Penal como consecuencia del `phishing´?". *UNED. Revista de Derecho Penal y Criminología, 3.ª Época, n.º 30,* pp. 273-304. 2023

[**5**] Feijoo Sánchez, Bernardo. "La teoría de la ignorancia deliberada en Derecho penal: una peligrosa doctrina jurisprudencial". *InDret. Revista para el Análisis del Derecho, nº 3*. 2015.

[**6**] Pérez Guerra, M. "Ciberdelitos y responsabilidad civil de las entidades financieras a la luz de la jurisprudencia", En *Revista de Derecho del Mercado de Valores, Nº 29,* 2021.

[**7**] Fernández Teruelo, J. "Clásicas y nuevas conductas fraudulentas ejecutadas en la red y su subsunción en los tipos de estafa y estafa informática contenidos en el código penal". En *Un modelo integral de derecho penal. Libro homenaje a la profesora Mirentxu Corcoy Bidasolo*, Agencia Estatal Boletín Oficial Del Estado, 2022.

[**8**] Ayerbe, A. "La ciberseguridad y su relación con la inteligencia artificial". *Análisis del Real Instituto Elcano (ARI),* Nº. 128, 2020.

Análisis y validación de un videojuego como herramienta de investigación para el ciberacoso

Andrea Baños Ramos[1], María Reneses Botija[2], Mario Castro Ponce[1], Edmond Awad[3,4], Gregorio López López[1]
[1]Instituto de Investigación Tecnológica (IIT), España
Escuela Técnica Superior de Ingeniería (ETSI), Universidad Pontificia Comillas, España
[2]Instituto de Investigación Tecnológica (IIT), España
Facultad de Ciencias Humanas y Sociales (CHS), Universidad Pontificia de Comillas, España
[3]Uehiro Oxford Institute, University of Oxford, UK
[4]Department of Economics, University of Exeter, Exeter, UK

abanos@comillas.edu, mreneses@comillas.edu, marioc@iit.comillas.edu, edmond.awad@uehiro.ox.ac.uk, gllopez@comillas.edu

***Resumen*—Este estudio explora el uso de un juego serio como una herramienta de investigación alternativa o complementaria a encuestas y cuestionarios validados para estudiar el ciberacoso. Se diseñó un videojuego basado en evidencia científica, incluyendo revisión de literatura, análisis de sentencias y entrevistas con acosadores, víctimas y expertos. Los participantes completaron cuestionarios previa y posteriormente al juego, y sus comportamientos en el juego se comparan con estos. Los resultados muestran que el videojuego estima la prevalencia del ciberacoso con cifras similares a otros estudios y aproxima mejor la relación entre el tiempo en Internet y la agresión. Por ello, se proponen los juegos serios como una nueva herramienta de investigación, destacando sus potenciales y limitaciones, y sugiriendo más investigación para poder validarlos.**

***Index Terms*—Ciberacoso, estudios de prevalencia, factores de riesgo, juegos serios, cuestionarios validados.**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

El ciberacoso es un fenómeno importante que afecta especialmente a niños y adolescentes en la actualidad. Poder cuantificar y comprender este problema es más complicado que el acoso tradicional, ya que éste no involucra interacciones personales cara a cara. Sin embargo, es fundamental para abordar el problema [1]. Actualmente, los estudios de prevalencia se utilizan para estimar el porcentaje de población afectada por este fenómeno. Por otra parte, las características sociodemográficas de los colectivos afectados se emplean como factores de riesgo, para poder diseñar campañas de prevención efectivas en escuelas europeas.

Los métodos tradicionales como encuestas o cuestionarios validados[1] son ampliamente utilizados para este tipo de estudios en ciencias sociales. Sin embargo, estos métodos pueden no estar adaptados correctamente para los menores[2], a pesar de los procesos de validez[3] y fiabilidad[4] [2]. En este caso, los juegos serios[5] [4] juegan un papel fundamental como una nueva herramienta de investigación. De esta manera, los menores toman las decisiones en contexto, asumiendo un rol en cada escenario del videojuego[6], tratando así, de mitigar el sesgo de observación que puede surgir cuando se sienten evaluados a través de un cuestionario validado [5].

En este artículo, se explora la posibilidad de utilizar los datos recopilados a través del juego serio desarrollado en el proyecto RAYUELA[7] [www.rayuela-h2020.eu] para estudiar el ciberacoso, como una alternativa a aquellos datos recogidos a través de encuestas o cuestionarios validados.

Este videojuego se ha diseñado basado en la evidencia científica, con una extensa y exhaustiva revisión de la literatura y el análisis de 46 sentencias europeas relacionadas con el ciberacoso. Además de realizar 33 entrevistas semiestructuradas a 8 agresores, 12 víctimas, y 13 expertos.

Se desarrollaron dos ciberaventuras relacionadas con el ciberacoso para medir escenarios de agresión o comportamientos que pueden considerarse ciberacoso, así como factores de riesgo, observando el comportamiento de los menores. Antes de obtener las credenciales para iniciar sesión en el videojuego, los menores deben rellenar un cuestionario para recopilar información acerca de su edad, género, lugar de nacimiento, antecedentes migratorios, tipo de escuela, orientación sexual, y el tiempo que pasan en Internet. Después de jugar al videojuego, deben rellenar otro cuestionario sobre si han experimentado situaciones de ciberacoso en el pasado, como víctima, agresor, o no han experimentado ninguna. No se guarda ningún tipo de información personal de los participantes y previamente firmaron un consentimiento informado adaptado a los menores (dependiendo de la edad y el país, algunos padres también firmaron este consentimiento).

II. Metodología

En este artículo, se examina la posibilidad de utilizar los escenarios del videojuego para medir conductas de agresión, con el fin de llevar a cabo estudios de prevalencia sobre la existencia de ciberacosadores comparado con los resultados proporcionados a través de un cuestionario validado. Por otra parte, se explora el efecto de medir factores de riesgo

[1]Evaluación de procedimientos, demostrando que el cuestionario mide el constructo que se busca obtener.
[2]Careciendo de un entorno cómodo, fluido e inmersivo.
[3]Examinando si ciertas aventuras del videojuego, miden el constructo que se quiere medir respecto al ciberacoso.
[4]Asegurando la consistencia de los elementos medidos.
[5]Juegos no diseñados con un carácter lúdico, si no con un fin formativo o como un medio de recopilación de datos para realizar diferentes estudios [3].
[6]Proporcionándoles un entorno no invasivo y amigable.
[7]El archivo ZIP que contiene el videojuego, se puede descargar en la pestaña RAYUELA GAME, download the game.

(concretamente, el tiempo que los menores pasan en Internet) fuera del juego (a través del cuestionario) y dentro del juego (a través de un escenario concreto del videojuego), comparándolo con conductas de ciberacoso medidas fuera del juego (con el cuestionario validado) y dentro del juego (a través de diferentes escenarios). Hasta donde los autores conocen, no hay trabajos previos que hayan abordado este problema en la literatura (comparando ambos, cuestionario validado y juego serio para tratar de medir el mismo constructo).

Las métricas estadísticas utilizadas para este análisis son análisis descriptivos flexplots, para ilustrar gráficamente la asociación entre variables y observar cómo cada variable se desvía de su valor esperado[8] (ver *Ec. (1)*) [6]; y tablas de contingencia bayesianas, para analizar la distribución e influencia de las asociaciones junto con el criterio del factor de Bayes, que representa la probabilidad relativa de los datos condicionado a la hipótesis alternativa (asociación entre variables) respecto a la probabilidad relativa de los datos condicionado a la hipótesis nula (independencia de las variables) (ver *Ec. (2)*) [7]. Estas tablas de contingencia bayesianas y el factor de Bayes se han calculado utilizando *JASP* [8].

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

donde O_{ij} representa el valor observado y E_{ij} indica el valor esperado.

$$BF_{10} = \frac{P(datos|H_1)}{P(datos|H_0)} \tag{2}$$

Estos términos representan la probabilidad de los datos dado que una de las hipótesis es cierta.

$$\begin{aligned} P(datos|H_1) &= \int p(s \mid H_1) L_i(s)\, ds \\ &= \frac{1}{\Delta s} \int_s^{s+\Delta s} e^{-\frac{\chi^2(s)}{2}}\, ds \end{aligned} \tag{3}$$

donde $p(s \mid H_1)$ es la probabilidad previa para la variable s[9], y $L_i(s)$ representa la verosimilitud para la variable s. Se sigue el mismo procedimiento para obtener $P(datos|H_0)$.

Tabla I: Escala de Jeffreys [9]

Factor de Bayes (H_1 vs H_0)	**Grados de evidencia**
1 to $10^{1/2}$	Apenas digno de mención
$10^{1/2}$ to 10	Sustancial
10 to $10^{3/2}$	Fuerte
$10^{3/2}$ to 10^2	Muy fuerte
$> 10^2$	Decisivo

Otras escalas para interpretar el factor de Bayes como la de Lee y Wagenmakers también son utilizadas en este tipo de estudios [10].

[8]Como todas las variables utilizadas para este análisis son categóricas, se emplean tests de chi-cuadrado para determinar la relación.

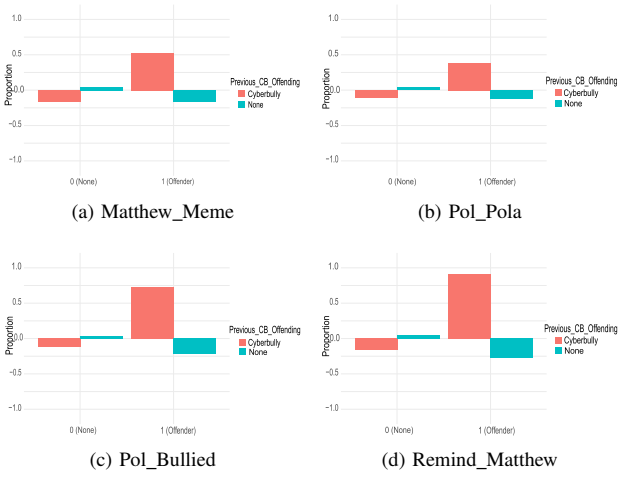
[9]Cualquier variable explicativa (sociodemográfica o pregunta del juego serio).

III. Análisis de resultados

El conjunto de datos utilizado para este análisis comprende una muestra de 826 individuos, compuesta por 495 niños, 298 niñas y 12 estudiantes de género no binario[10] de 11 países europeos (Bélgica, Croacia, República Checa, Estonia, Grecia, Hungría, Italia, Países Bajos, Portugal, España y Reino Unido) con edades comprendidas entre los 12 y 16 años.

Los escenarios relacionados con agresión[11] considerados en este análisis son la decisión de compartir un meme sobre un compañero de clase (`Matthew_Meme`), burlarse de la apariencia de un compañero (`Pol_Pola`) o ser testigo de una situación de acoso en la escuela (`Pol_Bullied`). Además, se considera un escenario donde se le pregunta al jugador sobre experiencias previas relacionadas con el ciberacoso (`Remind_Matthew`).

La *Fig. 1* muestra que el escenario `Remind_Matthew`, en el que los participantes deben autoinformar si han cometido ciberacoso en el pasado, presenta la mayor asociación con `Previous_CB_Offending`, seguido por el escenario de agresión `Pol_Bullied` [11].



(a) Matthew_Meme (b) Pol_Pola

(c) Pol_Bullied (d) Remind_Matthew

Figura 1: Flexplots. Asociación de Previous_CB_Offending con variables de agresión. Cuanto más grande sea la barra roja de la derecha, mayor será la influencia de esa variable en la identificación de menores en riesgo de cometer ciberacoso. Cuanto más baja sea la barra roja de la izquierda, menor será la cantidad de falsos negativos.

De acuerdo a numerosas revisiones, para medir ciberacoso es mejor utilizar una medida múltiple o agregada en vez de utilizar una medida individual, ya que captura mejor la complejidad y la variedad del fenómeno [12]. También proporciona una medida más objetiva de las experiencias de los jóvenes y reduce el riesgo de un posible bajo reporte de los casos [13]. Por lo tanto, se han considerado medidas agregadas de los escenarios de agresión, priorizando la especificidad (medida *AND*, habiendo actuado como agresor en todos los escenarios de agresión del videojuego) o la sensibilidad (medida *OR*, habiendo actuado como agresor en al menos un escenario

[10]21 estudiantes no contestaron la pregunta de género.

[11]Nuestros datos estarán disponibles de forma gratuita, por lo que incluimos aquí los nombres en clave utilizados en el videojuego para una referencia rápida.

de agresión). La *Tabla II* muestra las tablas de contingencia bayesianas y el factor de Bayes (*Tabla I*) para cada medida agregada comparándolas con `Previous_CB_Offending`.

Tabla II: Alta Prevalencia de Ciberacoso según la Medida CV: Tabla de Contingencia JS-CV (Variables Agregadas)

Prob. de Escenarios de Riesgo Agregados Cond. a Previous_CB_Offending	**Cyberbully**	**None**	**BF_{10}**
P(Agg_AND|CV=CB)	44.4 %	55.6 %	3.791
P(Agg_AND|CV=None)	21.9 %	78.1 %	
P(Agg_OR_rest|CV=CB)	38.1 %	61.9 %	613.413
P(Agg_OR_rest|CV=None)	20.1 %	79.9 %	
P(Agg_OR|CV=CB)	32.6 %	67.4 %	307338.028
P(Agg_OR|CV=None)	16.3 %	83.7 %	

Nota: BF_{10} (factor de Bayes) estadísticamente significativos. JS = preguntas del Juego Serio, CV = Cuestionario Validado, CB = Cyberbully (categoría de ciberacoso en el videojuego), Agg_AND = Agresión_AND (los escenarios Matthew_Meme, Pol_Pola y Pol_Bullied seleccionando la opción de "Agresión"), Agg_OR = Agresión_OR (los escenarios Matthew_Meme, Pol_Pola o Pol_Bullied seleccionando la opción de "Agresión"), Agg_OR_rest = Agresión_OR_restricted (los escenarios Matthew_Meme, Pol_Pola o Pol_Bullied seleccionando en al menos dos de ellos, la opción de "Agresión").

La *Fig. 2* muestra la prevalencia de individuos en riesgo de cometer ciberacoso en nuestro conjunto de datos, utilizando las medidas *AND*, *OR* y *OR_restricted* (habiendo actuado como agresor en dos o más escenarios de agresión). Se puede observar que, con las medidas *AND* y *OR_restricted*, se obtienen resultados de prevalencia en línea con otros estudios (20 % usando nuestro conjunto de datos y el cuestionario validado, 8,7 % en España y 14,9 % en Estonia según otro estudio basado en encuestas y muestras significativas realizado por nuestro equipo de investigación) [14].

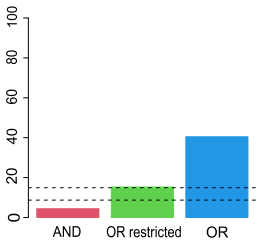


Figura 2: Prevalencia estimada obtenida al calcular la fracción de jugadores que actuaron como agresores en los tres escenarios del juego diseñados para ese propósito (*AND*, alta especificidad); en al menos dos de tres (*OR_restricted*, equilibrio entre sensibilidad y especificidad); o en al menos uno de los tres (*OR*, alta sensibilidad). Las líneas horizontales discontinuas representan la prevalencia estimada en otro estudio basado en encuestas y muestras significativas también realizado por nuestro equipo de investigación (8.7 % en España y 14.9 % en Estonia).

Sin embargo, la medida *OR* es empleada para analizar factores de riesgo, ya que el objetivo principal es la sensibilidad, tratando de no descartar ningún posible individuo en riesgo de cometer ciberacoso. La *Tabla III* muestra las tablas de contingencia bayesianas y el factor de Bayes (*Tabla I*) comparado con el tiempo que los menores pasan en Internet, medido dentro y fuera del videojuego con `Previous_CB_Offending` y la medida *OR*. Además, se observa que la mayor asociación aparece entre el factor de riesgo medido dentro del videojuego y la medida agregada *OR*, y que, en general, la asociación del factor de riesgo es más fuerte cuando se mide el comportamiento dentro del juego. Este análisis puede ser realizado con otros factores de riesgo como el apoyo familiar (donde se obtienen unos resultados similares).

Tabla III: Medición de Factores de Riesgo vs Medidas de Ciberacoso: Tabla de Contingencia JS-CV

Prob. del Tiempo Pasado En Línea Cond. al Ciberacoso	**CB_Off oG**	**CB_Off iG**
Factor de Riesgo oG	26.0 % [**7.323**]	34.5 % [0.199]
Factor de Riesgo iG	34.1 % [$\mathbf{6{,}544 \times 10^{6}}$]	52.9 % [$\mathbf{2{,}874 \times 10^{13}}$]

Nota: BF_{10} (aparecen entre corchetes) estadísticamente significativos. JS = preguntas del Juego Serio, CV = Cuestionario Validado, CB_Off = medida de Ciberacoso (oG = fuera del juego (CV), and iG = dentro del juego (JS)). CB_Off oG representsa la variable Previous_CB_Offending, CB_Off iG representa la variable Agresión_OR, Factor de Riesgo oG indica la variable Hours_Internet (CV), y Factor de Riesgo iG indica el escenario Time_Overrun (JS).

IV. Conclusiones

Este estudio trata de fomentar el debate sobre el uso de juegos serios como una herramienta de investigación y abrir nuevas líneas de investigación en esta dirección, como, por ejemplo, su uso para estudiar factores humanos en ciberseguridad. Los juegos serios presentan algunas características (p.ej., experiencias inmersivas o decisiones en contexto) que pueden ser beneficiosas para realizar ciertos estudios o estar dirigidos a determinados segmentos de la población, como niños y adolescentes. Sin embargo, también pueden generar algún tipo de error o sesgo, como cualquier otra medida o herramienta de investigación. Por lo tanto, se requiere una mayor investigación y debate al respecto.

Agradecimientos

Este proyecto de investigación ha sido posible gracias a la concesión de una beca estratégica por parte del Instituto de Investigación Tecnológica (IIT), como parte de la Universidad Pontificia de Comillas. Parte del trabajo también ha sido posible gracias a la financiación recibida por parte del equipo de investigación a través del proyecto europeo H2020 RAYUELA (contrato no. 882828) y del proyecto PID2022-140217NB-I00, financiado por MCIN/AEI/10.13039/501100011033 y por "ERDF A way of making Europe".

REFERENCIAS

[1] Chun, J., Lee, J., Kim, J., and Lee, S.: "An international systematic review of cyberbullying measurements", en *Computers in Human Behavior*, vol. 113, pp. 106485, 2020.

[2] Thomas, H. J., Connor, J. P., and Scott, J. G.: "Integrating traditional bullying and cyberbullying: Challenges of definition and measurement in adolescents–a review", en *Educational Psychology Review*, vol. 27, pp. 135-152, 2015.

[3] Morgan, C. B., Merrill, S. C., Clark, E. M., Wolfson, J. A., and Trubek, A. B.: "A serious games methodology to test solutions for regional food systems inequities", en *Journal of Rural Studies*, vol. 110, pp. 103366, 2024.

[4] Abt, C. C.: "Serious games" en *University press of America*, 1987.

[5] Freese, M., Lukosch, H., Wegener, J., and König, A.: "Serious games as research instruments–Do's and don'ts from a cross-case-analysis in transportation", en *European Journal of Transport and Infrastructure Research*, vol. 20, n. 4, pp. 103-126, 2020.

[6] Fife, D. A., Longo, G., Correll, M., and Tremoulet, P. D.: "A graph for every analysis: Mapping visuals onto common analyses using flexplot", en *Behavior Research Methods*, vol. 53, n. 5, pp. 1876-1894, 2021.

[7] Ly, A., Verhagen, J., and Wagenmakers, E. J.: "Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology", en *Journal of Mathematical Psychology*, vol. 72, pp. 19-32, 2016.

[8] Van Doorn, J., Van Den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, S., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., and Wagenmakers, E. J.: "The JASP guidelines for conducting and reporting a Bayesian analysis", en *Psychonomic Bulletin & Review*, vol. 28, pp. 813-826, 2021.

[9] Jeffreys, H.: "Theory of Probability", 3rd edition, Oxford University Press, London, pp. 95-103, 1961.

[10] Lee, M. D., and Wagenmakers, E. J.: "Bayesian Cognitive Modeling: A Practical Course", Cambridge University Press, 2014.

[11] Festl, R.: "Perpetrators on the internet: Analyzing individual and structural explanation factors of cyberbullying in school context", en *Computers in Human Behavior*, vol. 59, pp. 237-248, 2016.

[12] Sorrentino, A., Baldry, A., Farrington, D., and Blaya, C.: "Epidemiology of Cyberbullying across Europe: Differences between Countries and Genders", en *Educational Sciences: Theory & Practice*, vol. 19, n. 2, pp. 74-91, 2019.

[13] Betts, L. R., Gkimitzoudis, A., Spenser, K. A., and Baguley, T.: "Examining the roles young people fulfill in five types of cyber bullying", en *Journal of Social and Personal Relationships*, vol. 34, n. 7, pp. 1080-1098, 2017.

[14] Reneses, M., Parder, M. L., Riberas-Gutiérrez, M., and Bueno-Guerra, N.: "Cyberbullying and cyberhate as an overlapping phenomenon among adolescents in Estonia and Spain: Cross-cultural differences and common risk factors", en *Children and Youth Services Review*, pp. 108285, 2025.

Una revisión de: Leveraging Blockchain Usage to Enhance Slag Exchange

Aitor Gómez-Goiri*, Ivan Gutierrez-Aguero†, and David Garcia-Estevez‡
TECNALIA, Basque Research and Technology Alliance (BRTA)
Astondo Bidea 700, 48160 Derio, Biscay, Spain
Email: *aitor.gomez@tecnalia.com, †ivan.gutierrez@tecnalia.com, ‡david.garcia@tecnalia.com

***Resumen*—La escoria generada como subproducto de la fabricación de acero se puede usar para generar cemento, reduciendo residuos e impulsando la economía circular. Actualmente las empresas fabricantes de acero promueven acuerdos a largo plazo con una o unas pocas compañías cementeras a un precio fijo. En este trabajo proponemos una nueva solución que promueve reusar la escoria mediante su diferenciación a través de su agrupación por composición y un proceso de subasta. Este proceso se lleva a cabo en una red blockchain, aumentando la confianza en el sistema y facilitando garantías sobre la composición de la escoria a las compañías cementeras e indicadores de circularidad a reguladores.**

***Index Terms*—Circularidad, blockchain, cemento, acero, escoria, trazabilidad**

Tipo de contribución: *Investigación ya publicada*

I. Introducción

La industria siderúrgica genera como subproducto del proceso de fabricación del acero grandes volúmenes de escoria. Esta, a pesar de poseer propiedades valiosas para la producción de cemento, generalmente se trata como una materia prima homogénea mediante acuerdos bilaterales a largo plazo que no ponen en valor dichas escorias. En este contexto, el artículo "Leveraging Blockchain Usage to Enhance Slag Exchange" [1] presenta un novedoso modelo que aprovecha la tecnología blockchain para transformar el intercambio de escoria. La propuesta se enmarca en el esfuerzo por fomentar la economía circular y reducir la cantidad de residuos enviados a vertederos, en línea con las exigencias del Acuerdo Verde Europeo [2] y las oportunidades identificadas para el reciclaje de subproductos industriales.

El estudio destaca la importancia de diferenciar la escoria en función de su composición. La escoria puede contener compuestos beneficiosos como CaO, SiO_2 y Al_2O_3 que permiten a las empresas cementeras acceder a un producto de mayor calidad que optimiza el proceso de fabricación del cemento [3]. Por ello, se propone un sistema de subasta que, gestionado a través de una red blockchain [4], garantiza la trazabilidad, integridad y transparencia de la información compartida entre las partes involucradas, aprovechando las características inmutables y descentralizadas propias de esta tecnología [5].

II. Metodología y Diseño

La metodología aplicada identificó las deficiencias del modelo actual de intercambio de escoria y propuso un enfoque que mejora la eficiencia económica y refuerza la confianza entre los actores implicados. El modelo involucra a tres tipos de organizaciones: las empresas siderúrgicas, que generan diferentes tipos de escoria; las empresas cementeras, que requieren escoria de calidad para la producción de cemento; y las administraciones públicas, que desean promover la circularidad y minimizar el impacto ambiental.

El diseño del sistema se fundamenta en dos pilares principales. En primer lugar, se define un modelo de datos capaz de gestionar el ciclo de vida de la escoria, que abarca desde su generación en el proceso de refinación hasta su agrupación, almacenamiento, venta o eliminación. Este modelo contempla distintos estados de la escoria (individual, agrupada en lotes, almacenada y vendida) y permite registrar información crítica como la composición química y los volúmenes tratados. En segundo lugar, se implementa un mecanismo de subasta basado en contratos inteligentes, que posibilita a las empresas cementeras presentar ofertas por lotes de escoria, de modo que el proveedor (la empresa siderúrgica) pueda seleccionar la oferta que mejor se ajuste a sus criterios, ya sea en términos de precio, volumen o calidad.

III. Resultados

Para garantizar la integridad y trazabilidad de la información, se adopta la tecnología blockchain, la cual ofrece un entorno descentralizado y seguro [6]. Cada transacción, desde el registro inicial de la escoria hasta la aceptación de una oferta, se almacena en bloques interconectados, lo que impide la manipulación o alteración de los datos una vez registrados. La elección de una red permisionada, en este caso Hyperledger Fabric [7], permite limitar el acceso a los participantes autorizados y asegurar que la información sensible se maneje de forma privada, a la vez que se mantiene la transparencia en el proceso de subasta.

La implementación de la solución se realizó extendiendo el framework Traceblock, una plataforma de trazabilidad que facilita la creación y despliegue de contratos inteligentes para el seguimiento de activos en una red blockchain [8]. Este framework se adaptó para gestionar los distintos estados de la escoria, integrando operaciones específicas como el registro de composición, la agrupación en lotes, la presentación de ofertas y la aceptación o rechazo de las mismas. Asimismo, se desarrolló un API para permitir la interacción de los distintos actores a través de interfaces amigables, como aplicaciones de línea de comandos y dashboards web.

Durante la fase de validación, se comprobó que el sistema de subasta basado en blockchain mejora notablemente la transparencia y la precisión en la determinación de precios. Al disponer de información detallada sobre la composición y el volumen de escoria, las empresas cementeras pueden formular ofertas más ajustadas a sus necesidades, mientras que las siderúrgicas disponen de un mecanismo eficiente para

maximizar el valor obtenido por su subproducto [5]. Además, la implementación permitió la generación de indicadores de circularidad, facilitando la labor de las administraciones públicas en la elaboración y seguimiento de políticas ambientales y en el ajuste de tarifas de vertido [9].

Uno de los hallazgos clave fue que la utilización de contratos inteligentes y la inmutabilidad de la blockchain generan un entorno de confianza entre actores con intereses contrapuestos. Los datos registrados son accesibles para todos los participantes, lo que reduce la posibilidad de manipulación y fomenta la competencia leal en el mercado de la escoria. Por otro lado, se identificaron ciertos retos técnicos, como la necesidad de integrar mecanismos adicionales para preservar la privacidad de las transacciones y la posibilidad de ampliar la capacidad de procesamiento ante un incremento en el volumen de operaciones.

IV. Discusión

La solución propuesta evidenció el potencial de la tecnología blockchain para transformar sectores industriales tradicionales, en este caso, el mercado de la escoria. Al implementar un sistema de subasta transparente, se favorece la diferenciación del producto en función de su calidad, lo que puede incentivar a las empresas siderúrgicas a optimizar sus procesos de refinación y, a su vez, aumentar la reutilización del subproducto en la industria del cemento. Esta dinámica de mercado es esencial para la consecución de objetivos de economía circular, permitiendo reducir el volumen de residuos enviados a vertederos y minimizar el impacto ambiental derivado de la fabricación del acero [3].

No obstante, se reconoce la existencia de barreras en la adopción de soluciones blockchain en entornos industriales, como la necesidad de contar con una infraestructura tecnológica propia o la resistencia al cambio por parte de actores tradicionales [10]. Estas limitaciones, aunque importantes, pueden ser mitigadas mediante la implementación de redes permisionadas y estrategias de privacidad que aseguren que la información sensible no se vea comprometida, al mismo tiempo que se mantiene la transparencia del proceso de subasta [11]. Asimismo, el modelo propuesto podría extenderse a otras etapas de la cadena de suministro, abarcando desde la producción de cemento hasta el seguimiento del ciclo de vida de productos de construcción, lo que refuerza la integración de la trazabilidad en la economía circular [9].

Otro aspecto discutido es la escalabilidad de la solución. La elección de Hyperledger Fabric, con su capacidad para gestionar cientos de transacciones por segundo, demuestra que el sistema es viable incluso bajo escenarios de alta demanda, como la gestión de grandes volúmenes de escoria generados globalmente [7]. Sin embargo, se sugiere que futuras investigaciones evalúen el desempeño del sistema en condiciones reales de operación y comparen su eficiencia con otras plataformas blockchain para determinar la solución óptima en función de las necesidades específicas del sector.

V. Conclusiones

La integración de un sistema de subasta basado en blockchain para el intercambio de escoria representa una solución eficaz para fomentar la economía circular en la industria siderúrgica y cementera. El modelo propuesto mejora la transparencia y la equidad en la determinación de precios y facilita la obtención de indicadores de circularidad. La implementación mediante el framework Traceblock y la red Hyperledger Fabric ha demostrado que es posible establecer un entorno seguro y descentralizado, donde la confianza se refuerza mediante la inmutabilidad de los registros y la automatización de procesos a través de contratos inteligentes.

Aunque se identificaron desafíos en términos de privacidad, escalabilidad y la resistencia a la adopción, estos pueden ser abordados mediante futuras mejoras en la arquitectura y la incorporación de mecanismos adicionales de seguridad. En definitiva, la solución presentada abre la puerta a una nueva forma de gestionar el intercambio de subproductos industriales, potenciando el aprovechamiento de la escoria para la producción de cemento y contribuyendo a la reducción de residuos y a la sostenibilidad ambiental. Las líneas futuras de investigación incluyen la migración del sistema a redes públicas, el análisis comparativo de rendimiento y la ampliación del modelo para integrarlo en cadenas de suministro más amplias, consolidando así un ecosistema industrial basado en la transparencia y la eficiencia.

Agradecimientos

Esta investigación ha sido financiada por los proyectos de la Comisión Europea Horizonte 2020 «HyperCOG» (acuerdo de subvención nº 869886), e «ICEBERG» (acuerdo de subvención nº 869336).

Referencias

[1] A. Gómez-Goiri, I. Gutierrez-Aguero, and D. Garcia-Estevez, "Leveraging Blockchain Usage to Enhance Slag Exchange," *Applied Sciences*, vol. 14, no. 14, p. 6243, Jan. 2024, number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

[2] "The European Green Deal - European Commission." [Online]. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

[3] Q. Dong, G. Wang, X. Chen, J. Tan, and X. Gu, "Recycling of steel slag aggregate in portland cement concrete: An overview," *Journal of Cleaner Production*, vol. 282, p. 124447, Feb. 2021.

[4] A. Rejeb, A. Appolloni, K. Rejeb, H. Treiblmaier, M. Iranmanesh, and J. G. Keogh, "The role of blockchain technology in the transition toward the circular economy: Findings from a systematic literature review," *Resources, Conservation and Recycling Advances*, vol. 17, 2023, type: Review.

[5] Z. Shi, C. de Laat, P. Grosso, and Z. Zhao, "Integration of Blockchain and Auction Models: A Survey, Some Applications, and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 497–537, Jan. 2023.

[6] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized business review*, 2008.

[7] S. Shalaby, A. A. Abdellatif, A. Al-Ali, A. Mohamed, A. Erbad, and M. Guizani, "Performance Evaluation of Hyperledger Fabric," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, Feb. 2020, pp. 608–613.

[8] C. Regueiro, A. Gómez-Goiri, N. Pedrosa, C. Semertzidis, E. Iturbe, and J. Mansell, "Blockchain-based refurbishment certification system for enhancing the circular economy," *Blockchain: Research and Applications*, p. 100172, Nov. 2023.

[9] F. Corsini, N. M. Gusmerotti, and M. Frey, "Fostering the Circular Economy with Blockchain Technology: Insights from a Bibliometric Approach," *Circular Economy and Sustainability*, vol. 3, no. 4, pp. 1819–1839, Dec. 2023.

[10] H. Taherdoost, "A Critical Review of Blockchain Acceptance Models—Blockchain Technology Adoption Frameworks and Applications," *Computers*, vol. 11, no. 2, p. 24, Feb. 2022, number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[11] M. Arulprakash and R. Jebakumar, "Commit-reveal strategy to increase the transaction confidentiality in order to counter the issue of front running in blockchain," in *AIP Conference Proceedings*, vol. 2460, 2022, type: Conference paper.

Cooperación público-privada y protección de datos en la obtención transfronteriza de datos de localización como prueba electronica

Dr. Juan Carlos Ortiz-Pradillo*
Departamento de Derecho Procesal y Derecho Penal
Facultad de Derecho
Universidad Complutense de Madrid

juancarlosortiz@ucm.es

Dra. María Pilar Marco Francia
Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Derecho Procesal y Derecho Penal
Facultad de Derecho
Universidad Complutense de Madrid
Pilar.marco@ucm.es

***Resumen*- El artículo nos explica, dentro del margen jurídico al efecto, cómo la ciberseguridad también abarca el marco jurídico habilitante para que las autoridades españolas puedan obtener los datos de localización tratados y almacenados por empresas privadas (Google, Meta, Microsoft, etc.), con el fin de geolocalizar a los sospechosos, esto es, determinar la ubicación espaciotemporal de un sujeto en el lugar de los hechos delictivos. No obstante, sus fundamentos y conclusiones resultan extrapolables a la obtención de cualquier otro tipo de dato tráfico almacenado por terceros. De cara a implementar el marco jurídico de cooperación penal en materia de obtención transfronteriza de pruebas, la Unión Europea ha comprendido que la clave no gira en torno a la ubicación de la información, sino al origen de su creación: esos datos, con independencia de dónde se encuentren alojados, son fruto de conductas humanas llevadas a cabo en nuestro espacio económico y, por ende, deben resultar accesibles a las autoridades de nuestro espacio judicial.**

***Index Terms*- Datos, Localización, Pruebas, Ciberseguridad, Derecho Penal, Derecho Procesal**

Tipo de contribución: *Investigación original (límite 8 páginas).*

I. NATURALEZA Y MARCO JURÍDICO DE LOS DATOS DE LOCALIZACIÓN Y SERVICIOS DE VALOR AÑADIDO

El auge y generalización de las comunicaciones electrónicas ha conllevado la aparición de nuevas tipologías de datos generados con motivo de estas (*datos de tráfico*). Tal información puede resultar clave en la investigación de los delitos, motivo por el cual las autoridades siempre han prestado especial interés en obligar a las empresas a conservar dicha información y en obligarles a colaborar, con la cesión de los mismos, en la averiguación de los delitos (*data retention*).

Una nueva modalidad de esos datos de tráfico, fruto del progreso científico, lo representan los «datos de localización», que pueden ser resumidamente conceptualizados como una evolución o especificación (especie) de los datos de tráfico, e incluidos igualmente dentro de la categoría superior (género) de dato personal.

La Directiva sobre la privacidad y las comunicaciones electrónicas 2002/58/CE los define en el art. 2.c como *cualquier dato tratado en una red de comunicaciones electrónicas que indique la posición geográfica del equipo terminal de un usuario de un servicio de comunicaciones electrónicas disponible para el público* y, con más detalle, en su considerando núm. 14, que explica como tales datos *pueden referirse a la latitud, la longitud y la altitud del equipo terminal del usuario, a la dirección de la marcha, al nivel de precisión de la información de la localización, a la identificación de la célula de red en la que está localizado el equipo terminal en un determinado momento o a la hora en que la información de localización ha sido registrada*.

Su naturaleza como dato de tráfico resulta incuestionable: se alude a su condición como "dato de tráfico" en los considerandos núm. 15 y 35 de la misma. Pero también deben ser valorados como «servicios de valor añadido», definidos en el art. 2.g de la Directiva 2002/58/CE como todo servicio que requiere el tratamiento de datos de tráfico *o datos de localización distintos de los de tráfico que vayan más allá de lo necesario para la transmisión de una comunicación o su facturación*, entre cuyos ejemplos más habituales encontramos —considerandos núms. 18 y 35— servicios de "orientación vial, información sobre tráfico, previsiones meteorológicas o información turística".

En todo caso, estos datos —tráfico y localización— pertenecen a un género común: los «datos personales», pues así se declara expresamente en el art. 4 RGPD 2016/679, de 27 de abril de 2016 y en el correlativo art. 3 de la Directiva LED (*Law Enforcement Directive*) 2016/680, de igual fecha. Ello es importante porque, para su utilización como fuente de prueba en el proceso penal, las autoridades deberán cumplir las exigencias y garantías de tales normativas (licitud, limitación de la finalidad, calidad de la norma habilitante,

proporcionalidad, necesidad, etc[1].) según estemos ante unos datos objeto de tratamiento por las empresas y posteriormente cedidos a las autoridades estatales (RGPD) o ante una obtención y tratamiento por parte de las propias autoridades estatales para esos fines referidos al proceso penal (Directiva LED).

II. DATOS DE LOCALIZACIÓN Y ESCENARIO DEL CRIMEN

Los datos de geolocalización (datos de tráfico y de localización) de los dispositivos de los investigados se han convertido en una habitual fuente probatoria más con la que ubicarles en el escenario del crimen.

En los EE.UU., cada año aumentan exponencialmente las solicitudes policiales a empresas tecnológicas (ej., Google) de esos datos de localización, a través de una específica técnica de investigación criminal denominada «geofencing[2]»: con el objetivo de identificar a quien haya podido intervenir en una actividad criminal, las autoridades policiales delimitan un perímetro espacial de interés (por ej., las coordenadas geográficas del perímetro donde se encuentra la entidad bancaria en que se produjo un atraco), así como el lapso temporal en el que se calcula que se produjeron los hechos (por ej., los minutos previos y posteriore, al asalto). Y, a continuación, y mediante solicitud judicial (*geofence warrant*) se requiere a las empresas tecnológicas que indaguen en sus bases de datos y faciliten información detallada sobre los dispositivos a los que prestaron servicios dentro del referido perímetro espaciotemporal establecido, para de ese modo avanzar en las pesquisas e iniciar una línea de investigación centrada en las personas cuyos teléfonos móviles se se geolocalizaron en la escena del crimen.

III. COLABORACIÓN PÚBLICO-PRIVADA EN LA RECOLECCIÓN, CONSERVACIÓN Y USO DE DATOS PERSONALES

Las personas nos hemos convertido en la versión moderna de Hansel y Gretel y constantemente vamos dejando "migas de pan" en múltiples formatos digitales, a partir de las diversas interacciones electrónicas que llevamos a cabo a través de dichos dispositivos[3]. Esos datos son objeto de tratamiento y conservación por las empresas privadas, de modo que la actual tendencia legislativa a nivel mundial consiste en imponer a las empresas privadas una obligada «colaboración público-privada» consistente en específicos deberes legales de recolección, conservación y cesión de dichos datos a las autoridades encargadas de una investigación criminal.

El mayor ejemplo de ello se aprecia en la normativa referida a la obligación de conservación de datos de tráfico de comunicaciones electrónicas (Directiva 2006/24/CE y su desarrollo en España: la Ley 25/2007, de 18 de octubre, de conservación de datos relativos a las comunicaciones electrónicas y a las redes públicas de comunicaciones -en adelante, LCDCE-), pero hay muchos más ejemplos.

El Reglamento (UE) 2021/784 del Parlamento Europeo y del Consejo, de 29 de abril de 2021, sobre la lucha contra la difusión de contenidos terroristas en línea impone a los prestadores de servicios de alojamiento de datos el deber de conservar los contenidos terroristas que hayan sido retirados o a los cuales se haya bloqueado el acceso como consecuencia de una orden de retirada o de medidas específicas en virtud de los artículos 3 o 5 *así como cualesquiera datos conexos retirados como consecuencia de la retirada de dichos contenidos terroristas* que sean necesarios para la prevención, la detección, la investigación o el enjuiciamiento de delitos de terrorismo.

El Reglamento (UE) 2021/1232 del Parlamento Europeo y del Consejo de 14 de julio de 2021 por el que se establece una excepción temporal a determinadas disposiciones de la Directiva 2002/58/CE en lo que respecta al uso de tecnologías por proveedores de servicios de comunicaciones interpersonales independientes de la numeración para el tratamiento de datos personales y de otro tipo con fines de lucha contra los abusos sexuales de menores en línea, habilita a los proveedores de servicios de comunicaciones a emplear el uso de tecnologías específicas con el único fin de detectar y retirar el material de abuso sexual de menores en línea y denunciarlo a las autoridades policiales y a las organizaciones que actúen en interés público contra los abusos sexuales de menores, así como también incluye importantes obligaciones legales para dichos proveedores a la hora de almacenar los datos de contenido y los datos de tráfico conexos tratados para responder a las solicitudes emitidas por las autoridades y proporcionarles los datos necesarios para la prevención, detección, investigación o enjuiciamiento de infracciones penales establecidas en la Directiva 2011/93/UE.

Y el Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo de 19 de octubre de 2022 relativo a un mercado único de servicios digitales y por el que se modifica la Directiva 2000/31/CE (Reglamento de Servicios Digitales) también obliga a los prestadores de servicios de alojamiento de datos a notificar a las autoridades *cualquier sospecha de que se ha cometido, se está cometiendo o es probable que se cometa un delito que implique una amenaza para la vida o la seguridad de una o más personas*, habilitándose a las autoridades estatales a dirigirles «Órdenes de entrega de información» específica sobre uno o varios destinatarios individuales del servicio.

Ello da lugar a lo que BALKIN califica como «Estados Vigilantes»[4]: los Estados se sirven de las capacidades de las empresas privadas, no sólo para brindar servicios sociales valiosos, sino también para identificar problemas, evitar amenazas potenciales y combatir eficazmente la delincuencia.

Gracias a este régimen jurídico, las autoridades españolas pueden servirse de esos datos de tráfico y localización conservados por las operadoras de servicios de acceso a la telefonía móvil e internet para llevar a cabo dos eficaces diligencias de investigación: de una parte, reconstruir los movimientos del sospechoso y poder así conocer con gran precisión los lugares visitados, las rutas utilizadas, las personas con las que contactó, sus hábitos y rutinas, etc. (lo cual hemos

denominado como "geolocalización histórica[5]"), así como también identificar a sospechosos de actividades criminales que se encontraban en un determinado lugar porque sus teléfonos móviles recibían cobertura de una concreta antena BTS.

IV. DESLOCALIZACIÓN DE LA INFORMACIÓN Y OBTENCIÓN TRASNACIONAL DE DATOS COMO PRUEBA PENAL

Debido a que muchos de los servicios LBS de valor añadido, que dan lugar a la creación de bases de datos con «historiales de ubicación», están en poder de empresas a las que no resulta aplicable las obligaciones derivadas de la normativa sobre *Data retention* ni están sometidas a la jurisdicción española (ej., Google, Meta, Strava, Uber, Cabify, Bolt, Blablacar, etc.), las autoridades europeas no cuentan con la ventaja de las autoridades norteamericanas a la hora de servirse de esa diligencia de investigación calificada como «geofencing». Con razón, entre los años 2017 y 2018, el número de *geofence warrants* a Google se incrementó más de un 1.500%; entre 2018 y 2019 otro 500%, y se calcula que un 25% de las solicitudes policiales a Google se refieren a dichas solicitudes de geolocalización.

No obstante, existen instrumentos jurídicos internacionales para poder obtener esos codiciados datos de abonado, de tráfico y de localización.

1. Las órdenes de conservación y presentación de datos informáticos previstas en Convenios Internacionales

En materia de obtención transfronteriza de evidencias en formato digital, no solamente se demanda una mayor cooperación policial y judicial, sino que ésta sea cada vez más rápida y expeditiva a la hora de localizar, asegurar y remitir la información o fuentes de prueba reclamadas, dada su consabida volatilidad. Frente a los mecanismos convencionales de cooperación internacional —bilaterales o multilaterales— para la investigación y obtención de prueba penal, en donde rige el principio de reciprocidad o las tradicionales vías diplomática y de Autoridades Centrales designadas por los Estados parte, en la práctica se lleva a cabo una cooperación directa («direct request»), bien entre las autoridades policiales y judiciales de ambos países, bien entre aquéllas y los propios prestadores de servicios online.

Todo ello constituye, a nuestro parecer, una evidencia de la importancia de esa «colaboración público-privada» antes mencionada, conforme a la cual se asienta la cooperación directa entre el sector privado y las autoridades estatales en aras a acelerar la obtención transfronteriza de información electrónica. Esto se aprecia en el funcionamiento del instrumento jurídico más relevante —fuera del ámbito de la UE— como es el Convenio del Consejo de Europa sobre Ciberdelincuencia de 2001 (Convenio de Budapest), cuyos arts. 16, 17 y 18 prevén las órdenes de conservación rápida, revelación parcial y de presentación de datos almacenados por medio de un sistema informático, aplicables igualmente a la conservación y obtención de datos locativos.

El Departamento de Justicia estadounidense se ha visto desbordado ante el aluvión de peticiones de asistencia, debido a la que la información requerida se refiere a datos de cuentas de Internet correspondientes a esas *Big Tech* norteamericanas, por lo que viene aceptando —y supervisando— que las solicitudes de preservación se dirijan directamente a dichas compañías[6], para salvaguardar así la información, a la espera de que se emita y reciban la correspondiente comisión rogatoria por parte de la autoridad judicial.

Además, y como evolución de este sistema, tanto España como Estados Unidos han suscrito el segundo Protocolo adicional al Convenio de Budapest relativo al refuerzo de la cooperación y de la divulgación de pruebas electrónicas, abierto a la firma el 12 de mayo de 2022, y cuyo Preámbulo deja claro que uno de los motivos para su aprobación lo constituye *la necesidad de acceder a pruebas almacenadas electrónicamente en sistemas informáticos de jurisdicciones extranjeras y la necesidad de una mayor y más eficaz cooperación entre los Estados y el sector privado*, y que en este contexto *es necesaria una mayor claridad o seguridad jurídica para los proveedores de servicios y otras entidades en relación con las circunstancias en que pueden responder a las solicitudes directas de las autoridades de justicia penal de otras Partes para la divulgación de datos electrónicos*.

2. La Orden Europea de Investigación

La Orden Europea de Investigación (en adelante, OEI) constituye el instrumento más eficaz de cooperación en materia penal en el espacio judicial europeo, ideado con el objetivo de conseguir «la creación de un sistema general para obtener pruebas en los casos de dimensión transfronteriza, basado en el principio de reconocimiento mutuo» (Programa de Estocolmo), al resultar utilizable, tanto a los efectos de llevar a cabo una o varias medidas de investigación específicas en el Estado de ejecución de la OEI, como a los efectos de recabar fuentes de prueba que ya estén en posesión de la autoridad de ejecución. Y, además, porque dicha Directiva prevalece sobre los convenios internacionales de asistencia mutua, así como sobre los convenios celebrados en el marco del Consejo de Europa entre los Estados miembros, según afirma su considerando núm. 35.

En nuestra opinión, no deberían existir obstáculos a la hora de emitir una OEI para obtener los datos locativos almacenados por empresas ubicadas en el Estado de ejecución. Todos los Estados miembros cuentan con expresa legislación referida a la conservación y cesión de los datos de tráfico y de localización, y en su defecto, sobre obtención o secuestro de datos en formato digital almacenados por entidades públicas o privadas, pues todos los Estados han ratificado el Convenio de Budapest y cuentan con diligencias similares a la orden rápida de conservación de datos informáticos almacenados (art. 16) y la orden de presentación (art. 18), de modo que a la autoridad de emisión no debiera resultarle complicado especificar la diligencia solicitada (de acuerdo con lo dispuesto en el apartado e del art. 5.1 DOEI) y justificar su idoneidad, necesidad y proporcionalidad en función de los hechos objeto de investigación (art. 6.1 DOEI).

Es decir, difícilmente podrá la autoridad del Estado de ejecución acudir a una medida de investigación distinta por considerar que el «geofencing» no existe en su Derecho

nacional, con independencia del *nomen iuris* empleado, pues ya ha quedado aclarado que se trata de datos de conectividad generados por los dispositivos y que son recopilados, bien por parte de las empresas proveedoras de acceso a internet (ej., los datos CSLI), bien por parte de las empresas prestadoras de servicios de la sociedad de la información (ej., el historial de ubicaciones, también denominado en otras ocasiones como "actividad web y app" —*web and app activity, WAA*—).

3. Las «Órdenes e-evidence»: principio de efectividad y cooperación directa

A la hora de resolver el puzzle referido a la multiterritorialidad de la prueba en el ciberespacio, la UE ha dado un paso trascendental con la aprobación de las «Órdenes e-evidence»: la Orden Europea de Producción —EPOC, por sus siglas en inglés de *European Production Order Certificate*— y a la Orden Europea de Conservación —EPOC-PR, por sus siglas en inglés de *European Preservation Order Certificate*—, en virtud de lo recogido en el Reglamento (UE) 2023/1543 del Parlamento Europeo y del Consejo, de 12 de julio de 2023 (aplicable a partir del 18 de agosto de 2026).

Si el modelo iniciado por la OEI fue valorado muy positivamente, el futuro régimen europeo en materia de obtención transfronteriza de información electrónica instaurado a través del paquete *e-evidence* resultará mucho más ágil y puede llegar a convertirse en el paso definitivo para maximizar la eficacia de las investigaciones necesitadas de cooperación internacional, porque conduce a la europeización de la cooperación directa entre agentes estatales y empresas privadas (mayormente estadounidenses), es decir, institucionaliza la referida «colaboración público-privada» de la investigación criminal aquí examinada, por los siguientes motivos:

En primer lugar, los datos de localización constituyen expresamente una de las categorías de «datos de tráfico» que pueden ser objeto de tales órdenes EPOC y EPOC-PR (apartado 11 del art. 3). Con ello, el marco jurídico del paquete *e-Evicence* sería aplicable a la ejecución de la diligencia de investigación aquí calificada como «geofencing» para conseguir la obtención transfronteriza de dichos datos de localización que puedan haber sido objeto de tratamiento y conservación por prestadores de servicios que ofrezcan servicios de comunicaciones electrónicas o de valor añadido en la Unión Europea.

En segundo lugar, los destinatarios de dichas órdenes no son ya las autoridades del Estado de ejecución de la diligencia de investigación, sino *los prestadores que ofrezcan servicios en la Unión* y estén establecidos en ese otro Estado miembro o, no estándolo, estén representados por un representante legal en ese Estado miembro, *con independencia de la ubicación de los datos* (art. 1.1).

En tercer lugar, estimamos que resultan fáciles de cumplir las exigencias de contenido de una EPOC del artículo 5, como puedan ser la especificación por parte de la autoridad emisora de la categoría de los datos solicitados tal como se definen en el artículo 3, puntos 9 a 12 (art. 5.5.d); el período de tiempo que cubren los datos cuya producción se solicita (art. 5.5.e); los motivos para determinar que la orden europea de producción cumple las condiciones de necesidad y proporcionalidad establecidas en el apartado 2 (art. 5.5.i); o una descripción sucinta del caso (art. 5.5.j). Ya se ha especificado que esta diligencia de investigación consiste precisamente en que las autoridades investigadoras delimitan previamente las coordenadas geográficas de la zona, así como el concreto lapso temporal que consideren de interés para la averiguación de los hechos, a la hora de requerir a las operadoras que faciliten información que obre en sus bases de datos acerca de los dispositivos activos en dicho perímetro espacio-temporal, una vez ponderada la necesidad y proporcionalidad de dicha medida de investigación.

Y en cuarto lugar, porque la solicitud de los datos de localización referidos a un *geofencing* constituye una diligencia de investigación para la cual, inicialmente, no será necesario notificar la EPOC a la autoridad de ejecución, lo cual facilitará la agilidad de la respuesta directa por parte del prestador de servicios destinatario de dicha orden, y con ello, la aludida efectividad de la investigación pretendida, pues se reducirán las posibilidades de que la autoridad de ejecución alegue alguno de los motivos de denegación de la orden.

De conformidad con el art. 8.2, dicha notificación de la EPOC a la autoridad de ejecución no será precisa "si, en el momento de emitir la orden, la autoridad emisora tiene motivos razonables para suponer que: a) la infracción se ha cometido, se está cometiendo o es probable que se cometa en el Estado emisor, y b) la persona cuyos datos se solicitan reside en el Estado emisor". Es lo que el considerando 51 del Reglamento califica como «vínculos sustanciales y sólidos con el Estado emisor» a la hora de no exigirse ninguna notificación a la autoridad de ejecución.

El *geofencing* se caracteriza precisamente por ello: se trata de información electrónica localizada en otro Estado —o en servidores ubicados en distintos Estados—, pero referida a unos hechos cometidos físicamente en el Estado emisor (atraco a una sucursal bancaria, atentado terrorista o desórdenes públicos, agresión sexual, etc.) y cuyos datos solicitados (datos de localización de dispositivos activos en dichos lugares) se encuentran igualmente vinculados con el territorio del Estado emisor y se corresponderán con personas (los presuntos autores y, en su caso, las víctimas) sobre las que existen suficientes conexiones como para sostener que residen en dicho Estado a los efectos de permitirle acudir a este nuevo modelo de cooperación directa en la obtención de información digital.

V. CONCLUSIONES

El surgimiento de los datos de tráfico y de localización producidos como consecuencia de las interacciones digitales entre dispositivos ha abierto un nuevo escenario para las investigaciones penales en el que los Estados han comprendido que los gigantes tecnológicos juegan un rol esencial y conviene tenerlos como aliados a la hora de conseguir tan decisivas fuentes de prueba.

Ante tal poder amasado en manos privadas, se ha legislado un nuevo orden jurídico caracterizado por imponer a las empresas una «colaboración público-privada» consistente en específicos deberes de recolección, conservación y cesión de dichos datos a las autoridades encargadas de una investigación criminal, e incluso de denuncia de conductas ilícitas con motivo de la utilización de sus servicios, así como una «cooperación

directa» de las empresas con las autoridades estatales encargadas de las investigaciones penales.

De cara a implementar el marco jurídico de cooperación penal en materia de obtención transfronteriza de pruebas, la Unión Europea ha comprendido que la clave no gira en torno a la ubicación de la información, sino al origen de su creación: esos datos, con independencia de dónde se encuentren alojados, son fruto de conductas humanas llevadas a cabo en nuestro espacio económico y, por ende, deben resultar accesibles a las autoridades de nuestro espacio judicial. Dicho desde otra óptica, esa información electrónica genera ingentes beneficios económicos a las empresas de servicios digitales porque éstas hacen negocios en suelo europeo —es aquí desde donde recolectan esa preciada información que después monetizan—, independientemente de cuál sea su nacionalidad o de que tengan o no aquí sedes, personal o infraestructuras propias.

Al igual que ese «doing Business» y el concepto de «actividad dirigida» fueron manejados a la hora de fundamentar la licitud de los fueros de competencia judicial civil internacional en materia de contratación electrónica, esa información digital recolectada en suelo europeo permite sustentar también este paso adelante de la legislación europea en materia de obtención transfronteriza de la prueba electrónica con el paquete *e-evidence,* que va más lejos que la CLOUD Act, porque ésta se refiere a proveedores de servicios con sede estadounidense, mientras que el primero lo amplía a cualesquiera proveedores de servicios que hagan negocio en Europa, sin importar su sede principal.

Agradecimientos

El presente trabajo se ha realizado con los fondos del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation), a través de la Cátedra "Ciberseguridad para la Innovación y la Protección Digital" INCIBE-UCM.

Además, este trabajo ha contado con el apoyo de los siguientes Proyectos de Investigación:
Proyectos de la Comisión Europea: HEROES (acuerdo de subvención nº. 101021801) y ALUNA (https://aluna-isf.eu/), en virtud del acuerdo de subvención nº 101084929, IP Luis Javier García Villalba. El contenido de este artículo no refleja la opinión oficial de la Unión Europea ni de la Comisión Europea. Los autores son los únicos responsables de la información y las opiniones expresadas en él.

Proyecto I+D+i "PRUEBA PENAL Y NUEVOS RETOS EN EL PROCESO PENAL TRANSNACIONAL". Ref. PID2023-148413NB-I00, financiado por el Ministerio de Ciencia y Universidades, cuya Investigadora Principal es Lorena Bachmaier Winter.

Proyecto «Violencia extrema sobre la mujer: respuesta penal y alternativas para la prevención» (VEM, PID2022-140170NB-I00). 2022-2026. CO-IP's Carolina Bolea Bardon y Juan Carlos Hortal Ibarra. Entidad financiadora: Ministerio de Ciencia e Innovación del Gobierno de España.

Proyecto Criminalidad organizada transnacional y empresas multinacionales ante las vulneraciones a los derechos humanos. Nº: PID2020-117403RB-100. IP Laura Zúñiga Rodríguez. Programa financiador: Programa Estatal de I+D+I orientada a los Retos de la Sociedad, del Plan Estatal de Investigación Científica y Técnica y de Innovación de 2017-2020. Entidad financiadora: Ministerio de Ciencia e Innovación del Gobierno de España.

"Cuestiones controvertidas en relación con el ejercicio de la capacidad jurídica por los y las menores de edad" (Ref. CIGE/2023/71), financiado por la Conselleria de Educación, Universidades y Empleo de la Generalitat de Valencia. I.P. P. Chaparro Matamoros."

Referencias

[1] MONTORO SÁNCHEZ, J. A., "Los principios rectores del tratamiento de datos de carácter personal y sus implicaciones en el proceso penal", *Revista Acta Judicial*, Nº. 10, 2022, pp. 37-73.

[2] ORTIZ PRADILLO, J. C., "«Geofencing»: vestigios digitales, transnacionalidad de la prueba y defensa de los derechos fundamentales ante una nueva diligencia tecnológica de investigación", en *Prueba penal y derecho de defensa en la era digital: nuevos paradigmas y retos* (Coord. Lorena Bachmaier Winter), Aranzadi, 2024, p. 288.

[3] ORTIZ PRADILLO, J. C., "Big data, vigilancias policiales y geolocalización: nuevas dimensiones de los derechos fundamentales en el proceso penal", *Diario La Ley*, núm. 9955, 2021.

[4] BALKIN, J. M., «The Constitution in the National Surveillance State», *Minnesota Law Review 93* (1), 2008, pp. 1-25.

[5] ORTIZ PRADILLO, J. C., "Big data, vigilancias policiales y geolocalización: nuevas dimensiones de los derechos fundamentales en el proceso penal", *Diario La Ley*, núm. 9955, 2021. ATHI, M., "Rethinking Reverse Location Search Warrants", *The Journal of Criminal Law and Criminology*. 111 (3), 2021.

[6] VV.AA., "Geofence Warrants and the Fourth Amendment", *Harvard Law Review* 134 (7). May 10, 2021, pp. 2508-2529; RATHI, M., "Rethinking Reverse Location Search Warrants", *The Journal of Criminal Law and Criminology*. 111 (3), 2021, pp. 805-837.

Premios RENIC a la mejor tesis doctoral y al mejor trabajo fin de máster

Premio al mejor trabajo fin de máster

RENIC 2025 Mejor Trabajo Fin de Máster

TANUKIPOT: HONEYPOT BASADO EN GEMELOS DIGITALES PARA INFRAESTRUCTURAS CRÍTICAS

Máster en Ingeniería Informática
Especialidad en Ciberseguridad

Defensa: 07/03/2024



UNIVERSIDAD DE MÁLAGA

Alumno:
Javier Parada Tallante
Tutora:
María Cristina Alcaraz Tello
Cotutor:
Francisco Javier López Muñoz

Enlace documentos:
https://drive.proton.me/urls/Y9KJ8PWWZM#IwvT15CXviWQ

ISBN:978-84-10169-61-6

1 Enlace a documentos

El siguiente enlace incluye todos los documentos requeridos para la participación en RENIC para el premio de tipo MEJOR TRABAJO DE FIN DE MÁSTER EN CIBERSEGURIDAD. Incluye: documento del trabajo original y documentos que acreditan la calificación y fecha de la defensa. https://drive.proton.me/urls/Y9KJ8PWWZM#IwvT15CXviWQ

2 Novedad y aportación a la ciberseguridad

Este Trabajo de Final de Máster (TFM) surge de la necesidad emergente por parte de sectores críticos, tales como los ecosistemas industriales, por anticiparse a agentes maliciosos en un entorno cada vez más hostil y complejo. Estos sistemas industriales han vivido un cambio radical en la última década, formando sistemas cada vez mas complejos e interconectados. Este factor de potencial vulnerabilidad junto a la criticidad de las diversas industrias han hecho de este tipo de sector un objetivo atractivo para agentes y organizaciones maliciosas.

De hecho, el panorama actual de amenazas ha situado a las infraestructuras críticas en el punto mira [1], [2], lo que llama a su vez a la ciber-defensa activa o proactiva [3], siendo esta el **objetivo principal** del sistema propuesto en este TFM (*Tanukipot*), con el fin de comprender anticipadamente qué tipos de ataques son tendencia y así poder establecer una defensa preventiva, robusta y eficaz. Este enfoque proactivo no es arbitrario, es de especial importancia en este tipo de ecosistemas, ya que mecanismos convencionales como la detección de intrusiones, actúan una vez ya está sucediendo el ataque, o incluso, cuando este ya se ha realizado con éxito. Este tipo de defensa no es suficiente en estos entornos, haciendo de la defensa proactiva una necesidad en un entorno donde un fallo puede tener consecuencias económicas, sociales y humanas.

Una de las tecnologías existentes que típicamente ofrecen este enfoque proactivo a la defensa de infraestructuras críticas son los sistemas *honeypot*. Un *honeypot* es una herramienta utilizada en el campo de la ciberseguridad con el fin de detectar y anticipar diferentes tipos de ataque sobre sistemas y redes informáticos, sean conocidos o no. Esta herramienta simula un sistema real, ya sea total o parcialmente, con el objetivo de atraer, engañar y ser atacado por un agente malicioso. De esta forma, estos sistemas sirven como fuentes de datos para detectar y analizar comportamientos ofensivos, con el fin de desarrollar posteriormente: técnicas defensivas, reglas de detección, descubrimiento de vulnerabilidades, etc. Algunos esfuerzos por avanzar en esta dirección pueden verse en sistemas *honeypot* industriales ampliamente reconocidos como: *GridPot* [4] para redes eléctricas y *GasPot* [5] para redes oil y gas.

Con el fin de profundizar y aportar a la tecnología relativa a sistemas de *honeypot*, este

TFM pretende contribuir mediante el uso de gemelos digitales para una simulación realista. Estos sistemas actúan como una copia perfecta de una contraparte física de un sistema real. A diferencia de otros tipos de simulación, los gemelos digitales poseen un flujo de sincronización bidireccional entre la parte física y la simulada, permitiendo una mejora continua entre ambos. Esta tecnología no solo puede usarse como un sistema de predicción de determinados entornos, si no que sirve como elemento de engaño para atraer a *insiders* sin exponer el sistema real y su propiedad intelectual frente entidades externas. Este software también es idóneo para realizar pruebas con *malware*, con el objetivo de analizar su comportamiento e impacto en el sistema real, ya que no se trata de personas reales si no conjuntos de instrucciones maliciosas de las que se puede extraer información. Adicionalmente y dada la problemática que puede ocasionar determinados tipos de pruebas como de tipo *pentesting* sobre una infraestructura crítica real, los gemelos digitales con soporte de estrategias de engaño bajo controles en la comunicación bidireccional y fugas, se convierten también en una herramienta idónea para la prevención. Esto es debido a que permiten realizar pruebas de alta intensidad con una menor carga de limitaciones, al no estar atacando la infraestructura real sino a sus correspondientes copias virtuales ofreciendo un alto nivel de realismo al comportamiento del mundo físico que otras tecnologías no pueden aportar. Es por ello que este TFM también contempla un servicio adicional de *pentesting* automatizado. Este servicio, haciendo uso de técnicas de inteligencia artificial y fuentes de ciberinteligencia externas para conocer técnicas de explotación y vulneración, es capaz de atacar el sistema, aprendiendo de él y de sus vulnerabilidades. Esta funcionalidad integrada junto con el sistema *honeypot* conforma el principal aporte y novedad de este trabajo a la ciberseguridad, en el que se juega con una serie de tecnologías que en conjunto añaden estrategias de ataque inteligente y el descubrimiento preventivo de vulnerabilidades.

De forma resumida, **la contribución de *TanukiPot*** consiste en el despliegue de una copia fiel del sistema ciberfísico observado, basado en la tecnología de gemelos digitales, funcionando como *honeypot* y con el fin de utilizar dicha copia como el modo de aprender a su vez de las vulnerabilidades y debilidades del sistema real. Para potenciar el descubrimiento automático de vulnerabilidades, el *software* presentado cuenta con agentes inteligentes capaces de atacar, defender, e incluso, competir bajo los principios de teoría de juegos en el propio *honeypot*. Todo esto aprovechando la constante monitorización que aporta *TanukiPot* como *honeypot*, permitiendo analizar estos agentes inteligentes y sus acciones, y solucionar más tarde los fallos de seguridad del sistema real y dar garantías de continuidad. La herramienta ha sido probada en un escenario crítico en el que predomina las arquitecturas y protocolos de control industrial, que como se hacía referencia anteriormente, es de especial relevancia el despliegue de este tipo de técnicas y herramientas preventivas y defensivas.

3 Resultados obtenidos

El TFM presentado en este informe ha dado excelentes resultados tanto a nivel académico como a nivel formativo en la carrera del estudiante. En primer lugar, el TFM fue calificado con una nota de 10.0 y alcanzando la calificación de *Matrícula de Honor*, reflejando la calidad del trabajo al haber superado con la máxima calificación posible los criterios establecidos por la universidad.

En segundo lugar, el trabajo ha sido utilizado como base para la elaboración de un software más complejo y avanzado en el proyecto nacional *SecTwin 5.0* (TED2021-129830B-I00), financiado por el Ministerio de Ciencia e Innovación, la Unión Europea (Next Generation EU) bajo el Plan de Recuperación, Transformación y Resiliencia, y la Agencia Estatal de Investigación. De hecho, la aportación del TFM surge por los requisitos de difusión del proyecto en el que era requerido transferir conocimiento al mundo académico y a través de trabajos fin de Grado/Máster. Como un resultado, el trabajo no solo ha aportado valor académico intensificando las competencias de ciber-defensa, sino qué además, la solución (*TanukiPot*) está siendo integrada como parte de los procesos de defensa y prevención de la propia la Plataforma de SecTwin 5.0, capaz de atraer actores maliciosos, aprovechando el realismo presentado por su correspondiente gemelo digital (de la Plataforma).

En tercer lugar y al margen del reconocimiento académico, el software desarrollado como resultado de este trabajo, ha servido como base para la elaboración de un artículo académico, el cual será enviado (antes del 22/4/2025) a la conferencia *European Symposium on Research in Computer Security (ESORICS 2025)* [6], una de las conferencias de ciberseguridad más relevantes a nivel europeo e internacional (ICORE: CORE A [7]). En este caso, y a diferencia del proyecto *SecTwin 5.0*, se ha utilizado las capacidades de simulación del sistema para ejecutar múltiples ataques con sistemas de inteligencia artificial y auto aprendizaje, a fin de intensificar la ofensiva y la inteligencia del proceso de *pentesting*.

Por último, este TFM ha marcado la línea de investigación del estudiante en su actual Tesis Doctoral, especializándose en *Cyber Threat Hunting (CTH)* o caza de ciber amenazas, y, más concretamente, habiendo profundizado en el campo de la emulación adversaria a lo largo de su primer año.

4 Resumen

Este TFM presenta *TanukiPot*, un *honeypot* capaz de representar una imagen exacta de un entorno real crítico y monitorizarlo en busca de amenazas. *TanukiPot* está formado por cuatro componentes principales: los gemelos digitales, el sistema *honeypot*, los agentes inteligentes y el sistema de auditoría de los cuales se trata en mayor profundidad en los siguientes apartados.

4.1 Gemelos Digitales

Los gemelos digitales son la pieza clave de este software, puesto que son el objeto de ataque de los agentes maliciosos y el sujeto de pruebas y estudio para establecer mejoras en la contraparte física del sistema de forma automatizada. Los gemelos digitales son una representación digital exacta de una contraparte física, con la sutil diferencia de que existe un flujo bidireccional de información de sincronización de la que ambas partes se retroalimentan. En el caso de los gemelos digitales desplegados por *TanukiPot*, reciben información de sincronización de su contraparte física para estar actualizados en todo momento del estado actual del sistema. Este estado de sincronización incluye: cambios en el *software* o en el *hardware* de los componentes, aparición de nuevos dispositivos que se conectan a la red, eliminación de estos dispositivos, cambios en la topología de la red etc. Por otra parte, *TanukiPot* permite una conexión inversa desde el sistema del gemelo digital en dirección a su contraparte física, donde este informa de vulnerabilidades encontradas para que sean reparadas, debilidades del sistema en la arquitectura de red, información de la debilidad de cada dispositivo etc.

Adicionalmente, *TanukiPot* no solo permite la ejecución de un gemelo digital por cada contraparte física, si no que permite la creación de numerosas instancias digitales de una sola infraestructura crítica, manteniendo todas ellas sincronizadas a la vez. Esto se realiza de forma escalable, mediante un Application Programming Interface (API) basada en microservicios, donde cada acción y cada ejecutable puede desplegarse de forma independiente como una aplicación y comunicarse de forma asíncrona. Adicionalmente, la aplicación se estructura siguiendo los principios de Arquitectura Hexagonal [8], lo que proporciona modularidad a la aplicación y a la implementación utilizada permitiendo cambiarse fácilmente.

Por último, la implementación de virtualización realizada se sustenta sobre contenedores *Docker*, debido a su especial eficiencia sin perder aislamiento entre los diferentes procesos del sistema. Cabe añadir, que gracias a la modularidad que aporta la Arquitectura Hexagonal citada anteriormente, la infraestructura puede cambiarse, o incluso, combinarse, haciendo uso, por ejemplo, de máquinas virtuales, o incluso, un híbrido de máquinas virtuales y contenedores.

4.2 HoneyPot

Esta sección trata el desarrollo del sistema *honeypot*, un sistema que se ejecuta como un capa paralela a la ejecución de los gemelos digitales. Aunque como se cita en la sección anterior, la infraestructura del sistema de virtualización es modular, este apartado se centra en la implementación sobre contenedores *Docker* que se ha desarrollado para este TFM.

Para la monitorización de los diferentes gemelos digitales se ha desarrollado la herramienta *Tanuki Sniffer*. Esta herramienta es un analizador de paquetes, es decir, una herramienta capaz de detectar todos los paquetes de red enviados y recibidos por un dispositivo a través de todas sus interfaces. Esta herramienta trabaja a nivel de contenedores *Docker*, sin embargo, lo hace de forma particular. En lugar de ser un proceso en ejecución en el contenedor *Docker* que realiza un esnifado de paquetes, este *software* ejecuta una instancia de un contenedor paralelo con el que comparte su pila de red. De esta forma, *Tanuki Sniffer* es capaz de detectar todos los paquetes enviados y recibidos por un nodo del gemelo digital sin perturbar su ejecución.

Las ventajas que este sistema tiene son: en primer lugar, el realismo de la ejecución, no creando ningún proceso adicional dentro del proceso del gemelo digital que envíe y reciba información ni tenga un consumo de CPU. Esto es especialmente importante, ya que la ejecución de un analizador de paquetes en el *honeypot* podría resultar en una pista muy útil para un atacante de que está en un entorno no real de pruebas. En segundo lugar, un analizador de paquetes interno puede falsear las métricas de uso de recursos del sistema, consumiendo uso de disco, CPU etc. que puede ser confundido con el de un atacante.

En resumen, el sistema *honeypot* de *TanukiPot* es un sistema independiente de la ejecución de los gemelos digitales, actuando a modo de monitor de todas las actividades de red que suceden en todos los dispositivos y en todas las conexiones de estos. Este sistema se ejecuta de forma aislada de los procesos del gemelo digital a excepción de la pila de red, recurso que comparten, siendo indetectable para un atacante que ya haya vulnerado el sistema, pero proporcionando toda la información al *honeypot*.

4.3 Agentes Inteligentes

Los agentes inteligentes son un componente adicional de *Tanukipot* para agilizar la detección de vulnerabilidades y debilidades del sistema. Además, gracias a los informes que estos agentes generan, con dicha información se cierra el ciclo bidireccional de información con los gemelos digitales. La información generada por estos sistemas es clave para la subsanación de forma automatizada de la contraparte física sin necesidad de estar a la espera de intrusiones reales.

Los agentes inteligentes constan de dos partes, el primer componente consiste en un escáner y explotador de vulnerabilidades llamado *Tanuki Vuln Analyzer*, el cual, haciendo uso de una base de datos actualizada de vulnerabilidades, es capaz de analizar y explotar sistemas de forma genérica. La base de datos utilizada por este analizador es *Tanuki Vuln Guardian*, otro componente de la aplicación que presta información actualizada de ciberinteligencia de diferentes fuentes, como: *MITRE* y *ExploitDB*. El vector de ataque desarrollado por estos dos componentes en conjunto consta de las siguientes tres fases:

1. Reconocimiento: el analizador ejecuta un análisis de los puertos de un sistema. De este reconocimiento, obtiene información relevante como: puertos abiertos del sistema, sistema operativo, software que se esta ejecutando en cada puerto junto con su versión, etc.

2. Análisis: este proceso consiste en cruzar la información del software y componentes detectados con la base de datos de vulnerabilidades y *exploits*. Una vez se ha hecho este análisis, se preparan los *exploits* y *payloads* a ejecutar en un orden y frecuencia determinados.

3. Explotación: esta fase consiste en la ejecución concatenada de *exploits*. El objetivo final de esta fase es conseguir una *reverse shell* con permisos de administrador desde la cual poder propagarse a otros dispositivos de la red.

La segunda parte que conforman los agentes inteligentes es la inteligencia a nivel de red. Siguiendo con la Arquitectura Hexagonal, el sistema hace una abstracción de los algoritmos que cada agente puede ejecutar. *Tanukipot* proporciona un espacio de memoria para cada agente, donde poder guardar modelos de inteligencia artificial o algoritmos basados en experiencia. De esta forma, a base de ensayo y error pueden desarrollarse diferentes estrategias mediante algoritmos de aprendizaje que decidan que nodo atacar en cada momento. Adicionalmente, el sistema proporciona la capacidad de realizar equipos, tanto atacantes (con competencia entre ellos por vulnerar un sistema) como defensivos, cuyo objetivo sea neutralizar a los atacantes. Esta funcionalidad permite ejecutar aprendizaje adversario y hacer estudios bajo teoría de juegos para probar diferentes estrategias.

En concreto y para la elaboración de este TFM, se ha realizado una implementación de un agente inteligente basado en aprendizaje por refuerzo. Este tipo de agente inteligente ha sido especialmente eficaz, llegando a vulnerar sistemas de mas de 50 dispositivos conectados con un ratio superior al 90% de éxito. Estos resultados son de hecho los expuestos en la sección 3 como parte de la publicación académica citada.

4.4 Sistema de auditoría

El último componente de *Tanukipot* es el sistema de auditoría. Este sistema consiste un catalogo de software *Grafana*, *Loki* y *Promtail* que, capturando todos los *logs* de la salida estándar de los contenedores *Docker*, es capaz de crear un Panel de administración y monitorización del sistema. Esto incluye no solo a los gemelos digitales y todos sus *logs*, si no que además incluye todos los sucesos que ocurren en la plataforma *TanukiPot*, su base de datos y el entrenamiento de los agentes inteligentes. Este sistema es especialmente útil para monitorizar los componentes, las ejecuciones de diferentes *software* en el gemelo digital y el correcto entrenamiento de los agentes inteligentes en largos intervalos de tiempo.

Referencias

[1] Agencia de la Unión Europea para la Ciberseguridad (ENISA), *Ciberseguridad de Sectores Críticos*, Accedido el 15 de abril de 2025, 2025. dirección: https://www.enisa.europa.eu/topics/cybersecurity-of-critical-sectors.

[2] European Union Agency for Cybersecurity (ENISA), *Speech at AECA Round Table*, Accedido el 15 de abril de 2025, 2017. dirección: https://www.enisa.europa.eu/sites/default/files/all_files/2017-20-11%20ED%20speech%20at%20AECA%20round%20table.pdf.

[3] European Union Agency for Cybersecurity (ENISA), *Proactive Detection – Good Practices Gap Analysis Recommendations*, Accedido el 15 de abril de 2025, 2020. dirección: https://www.enisa.europa.eu/publications/proactive-detection-good-practices-gap-analysis-recommendations.

[4] sk4ld, *GridPot: Open source tools for realistic-behaving electric grid honeynets*, https://github.com/sk4ld/gridpot, Accessed: 2025-04-14, 2025.

[5] S. Hilt, *GasPot: Honeypot simulating a Veeder Root Guardian AST*, https://github.com/sjhilt/GasPot, Released at Black Hat 2015. Accessed: 2025-04-14, 2015.

[6] ESORICS 2025 Program Committee, "30th European Symposium on Research in Computer Security (ESORICS 2025)," en *Proceedings of the 30th European Symposium on Research in Computer Security (ESORICS 2025)*, Conference dates: September 22–26, 2025, Toulouse, France, sep. de 2025. dirección: https://esorics2025.sciencesconf.org/.

[7] Computing Research and Education Association of Australasia (CORE), *European Symposium on Research in Computer Security (ESORICS)*, Clasificación CORE 2023: A. Campo de investigación: 4604 – Ciberseguridad y privacidad. Accedido el 15 de abril de 2025, 2023. dirección: https://portal.core.edu.au/conf-ranks/515/.

[8] M. Cambarieri, F. Difabio y N. García Martínez, "Implementación de una Arquitectura de Software guiada por el Dominio," en *XXI Simposio Argentino de Ingeniería de Software (ASSE 2020)-JAIIO 49 (Modalidad virtual)*, 2020.

Premio a la mejor tesis doctoral

VII Premios de Investigación en Ciberseguridad

Mejor Tesis Doctoral en Ciberseguridad

Red de Excelencia Nacional de Investigación en Ciberseguridad (RENIC)

- **Título**: *Advances in Cybersecurity Incident Prevention and Analysis*
- **Autor**: Uroz Hinarejos, Daniel
- **Universidad**: Universidad de Zaragoza
- **Fecha de Lectura**: 19/06/2024
- **Menciones**: *Cum Laude*, Doctorado Internacional
- **Director**: Rodríguez Fernández, Ricardo Julio
- **Acceso**: https://www.educacion.gob.es/teseo/mostrarRef.do?ref=2528481 (bajo embargo durante un año, disponible el acceso a través de un correo a duroz@unizar.es)

1. Novedad del trabajo y aportación a la ciberseguridad

En el panorama digital actual, las organizaciones –independientemente de su tamaño– son susceptibles de experimentar incidentes de ciberseguridad durante el desarrollo de su actividad, por lo que es primordial que implementen estrategias sólidas tanto para anticiparse como para responder eficazmente a tales amenazas. Estos incidentes se manifiestan de muchas formas, cada cual con su propia intención, capacidad y técnica, lo que requiere de avances en las medidas de seguridad que no sólo minimicen la exposición a nuevas amenazas, sino que también analicen y mitiguen rápidamente los potenciales riesgos. Dada la complejidad de estos incidentes, las organizaciones deben dar prioridad a la implantación de un plan de respuesta a incidentes bien estructurado para gestionar y recuperarse eficazmente de las brechas de seguridad.

Para hacer frente a la naturaleza heterogénea de los incidentes de ciberseguridad, las organizaciones se basan en un proceso integral de respuesta a incidentes que incluye cuatro fases clave: (i) *preparación*; (ii) *detección y análisis*; (iii) *contención, remediación y recuperación*; y (iv) *actividad posterior al incidente*. Al mismo tiempo, los adversarios siguen un meticuloso proceso para ejecutar sus intrusiones, en el que cuentan con su propio conjunto de tácticas para lograr sus objetivos. Comprender estas tácticas de los adversarios es fundamental para que las organizaciones puedan defenderse de forma proactiva y mitigar las posibles ciberamenazas.

La tesis doctoral explora tres temas de la respuesta a incidentes: prevención de incidentes, análisis preliminar de incidentes y análisis en profundidad de incidentes. En primer lugar, presentamos un análisis comparativo y una evaluación de los protocolos IoT para la exfiltración de datos. En segundo lugar, mostramos

ISBN:978-84-10169-61-6

cómo el análisis forense de memoria puede ayudar a identificar rápidamente la presencia de malware o descartar procesos legítimos en un triaje inicial. Por último, exploramos cómo el análisis dinámico y la ejecución simbólica pueden ayudar a analizar ejecutables que implementan protocolos de red desconocidos.

Nuestras principales conclusiones son que los protocolos IoT MQTT y AMQP emergen como las opciones más adecuadas para la exfiltración de datos, debido a su optimización para la transferencia de datos y a su uso común en conexiones IoT en la nube a través de puertas de enlace. También introducimos una taxonomía de los métodos de persistencia de Windows y destacamos su importancia en el análisis forense de memoria. Además, exploramos la eficacia de basarse en la firma de código en el análisis forense de memoria, destacando su limitada utilidad debido a la presencia de datos incompletos y cambios causados por el tiempo de ejecución del sistema. Además, introducimos un sistema que utiliza gráficos de flujo de control dinámicos contaminados y ejecución simbólica para analizar binarios que implementan protocolos de red desconocidos; aunque esta propuesta puede ayudar a un analista, actualmente se enfrenta a limitaciones como la dependencia de códigos de estado en el protocolo, requisitos para el conjunto válido de datos iniciales y la visión parcial de la comunicación con protocolos que implican múltiples conexiones de *socket*.

2. Resultados del trabajo

Se han publicado tres artículos en revistas revisadas por pares y uno más está actualmente en proceso de revisión en otra revista revisada por pares. También se publicó otro durante el programa de doctorado, pero no se ha incluido en esta tesis.

Artículos de revista

[1] **D. Uroz** and R. J. Rodríguez. "Characteristics and Detectability of Windows Auto-Start Extensibility Points in Memory Forensics," in *Digital Investigation*, vol. 28, pp. S95-S104, Apr. 2019. DOI: 10.1016/j.diin.2019.01.026. Impact factor (JCR): 1.938, rank 108/156 (Q3).

[2] **D. Uroz** and R. J. Rodríguez. "On Challenges in Verifying Trusted Executable Files in Memory Forensics," in *Forensic Science International: Digital Investigation*, vol. 32, p. 300917, Apr. 2020. DOI: 10.1016/j.fsidi.2020.300917. Impact factor (JCR): 2.406, rank 119/223 (Q3).

[3] **D. Uroz** and R. J. Rodríguez. "Characterization and Evaluation of IoT Protocols for Data Exfiltration," in *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19062-19072, Oct. 2022. DOI: 10.1109/JIOT.2022.3163469. Impact factor (JCR): 11.1, rank 5/251 (Q1).

Bajo revisión

[4] **D. Uroz**, R. J. Rodríguez, S. Bardin, and F. Recoules. "Back to the Future: Uncovering Network Protocol States through Symbolic Execution and Taint Analysis," in Computer & Security. Impact factor (JCR): 5.6, rank 41/158 (Q2).

Otras publicaciones no incluídas en la tesis

[5] D. Uroz and R. J. Rodríguez. "Evaluation of the Executional Power in Windows using Return Oriented Programming," in *2021 IEEE Security and Privacy Workshops (SPW)*, San Francisco, CA, USA, 2021, pp. 361-372, DOI: 10.1109/SPW53761.2021.00056. GGS Class: Not indexed.

Proyectos de código abierto

En aras de la ciencia y la reproducibilidad, cuatro proyectos se han puesto en código abierto bajo distintas licencias públicas generales de GNU:

- WINESAP: Un plugin para Volatility para buscar todos los *Autostart Extensibility Points (AESPs)*, el subconjunto de puntos de extensibilidad del sistema operativo y de las aplicaciones que permiten que un programa se inicie automáticamente sin invocación explícita del usuario. Disponible bajo licencia GNU AGPLv3 en https://github.com/reverseame/winesap (accedido el 21 de abril de 2025).
- SIGCHECK: Un plugin de Volatility para verificar las firmas digitales de archivos ejecutables (es decir, archivos `.exe`, `.dll` y `.sys`) en volcados de memoria. Disponible bajo licencia GPLv3 en https://github.com/reverseame/sigcheck (accedido el 21 de abril de 2025).
- CHITON: Una biblioteca de Python para filtrar datos encapsulándolos en paquetes de protocolos IoT. Disponible bajo licencia GNU GPLv3 en https://github.com/reverseame/chiton (accedido el 21 de abril de 2025).
- SYNOPTIC: Proyecto cuyo objetivo es deducir el formato de los mensajes de protocolo y la máquina de estados finitos (FSM) de un protocolo desconocido. Disponible bajo licencia GNU GPLv3 en https://github.com/reverseame/synoptic (accedido el 21 de abril de 2025).

3. Resumen de la Tesis Doctoral

Esta tesis doctoral ha centrado la investigación en cómo las organizaciones pueden prepararse mejor ante incidentes de ciberseguridad, cómo detectarlos rápidamente y cómo analizarlos en profundidad cuando suceden. Para ello, esta tesis doctoral se estructura en torno a tres pilares fundamentales de la respuesta a incidentes: prevención, análisis preliminar y análisis en profundidad; ofreciendo contribuciones en cada una de estas fases.

Motivación y preguntas de investigación

En el mundo digital actual, en el que las organizaciones también forman parte, todos los sistemas están expuestos a sufrir incidentes, sin importar el tamaño de la organización. De igual forma que las defensas han evolucionado, es una realidad que los adversarios también han evolucionado para estar a la par. En este contexto, la tesis doctoral se plantea cuatro preguntas de investigación:

- ¿Podemos prepararnos ante adversarios que usan nuevos medios de comunicación durante un incidente?

- ¿Podemos detectar procesos sospechosos en un sistema comprometido rápidamente?
- ¿Podemos descartar rápidamente procesos legítimos?
- ¿Podemos mejorar el análisis de ejecutables sospechosos que se comunican con protocolos desconocidos?

Estas preguntas guían los tres bloques principales de la tesis.

Parte I: Prevención de incidentes mediante protocolos IoT

En esta parte se estudió cómo los protocolos de Internet de las Cosas (IoT) pueden ser utilizados de forma maliciosa para exfiltrar información desde una red comprometida. Se analizaron tres protocolos de IoT: CoAP, MQTT y AMQP, desde una perspectiva de eficiencia en la transmisión de información, la sobrecarga y su capacidad para evadir la detección.

Para evaluar esta hipótesis, se desarrollo Chiton, una herramienta que encapsula datos en mensajes de protocolos IoT. Los resultados de evaluación muestran que MQTT y AMQP son especialmente eficaces para la exfiltración, no solo por su eficiencia, sino también por su presencia común en redes empresariales legítimas. Además, se extendió la biblioteca Scapy para soportar AMQP y se experimentó entre un cliente (Raspberry Pi 3) y un servidor remoto para medir tiempos de transmisión y validación de integridad.

Los resultados muestran que los protocolos MQTT y AMQP superan claramente a CoAP en tiempo de exfiltración y robustez, lo cual plantea un nuevo vector de ataque subestimado en redes IoT.

Parte II: Análisis preliminar con forense de memoria

En los primeros minutos de una respuesta a incidente, la rapidez en la toma de decisiones es crítica. Por ello, se plantearon dos enfoques del análisis forense de memoria para detectar *malware*. Ambos enfoques son complementarios y ayudan a realizar un triaje más eficaz durante los primeros pasos del análisis de un incidente.

Persistencia mediante puntos de autoarranque (ASEPs) Se desarrolló Winesap, un plugin de Volatility que detecta claves de registro sospechosas que permiten la persistencia de malware en sistemas Windows. Se propuso una taxonomía de los ASEPs dividida en cuatro categorías: mecanismos del sistema, abuso del cargador de programas, abuso de aplicaciones y abuso del comportamiento del sistema.

Winesap identifica comandos y rutas sospechosas, flujos alternativos de datos NTFS y ejecuciones indirectas utilizando `rundll32`. Los experimentos demostraron su capacidad para localizar claves maliciosas con alta efectividad.

Verificación de ejecutables legítimos mediante firma de código Se desarrolló Sigcheck, otro plugin para Volatility, que analiza ejecutables en memoria y verifica si están firmados con certificados válidos (verificando las firmas de *Microsoft Authenticode*). El objetivo es permitir al analista centrarse

en procesos no firmados o potencialmente alterados. Aunque útil, se encontraron varias limitaciones: muchos procesos legítimos no están firmados y otros firmados pueden estar manipulados o tener firmas revocadas. Además, la paginación de memoria puede limitar la capacidad de verificación de los procesos.

Parte III: Análisis en profundidad de protocolos desconocidos

En la última parte, se aborda el análisis avanzado de binarios sospechosos que utilizan protocolos de red desconocidos o personalizados. Ante la dificultad de analizar estas comunicaciones, se propone la herramienta SYNOPTIC, que combina ejecución simbólica, análisis de flujo de control dinámico contaminado (*taint analysis*), y generación iterativa de mensajes de red.

La idea es inferir automáticamente la gramática del protocolo usado por el ejecutable y descubrir nuevos mensajes válidos. Como prueba de concepto, se avaluó SYNOPTIC sobre una implementación real de FTP en ProFTPD, obteniendo diagramas de estados útiles para entender su comportamiento.

Aunque útil, SYNOPTIC tiene limitaciones como la dependencia de la implementación de códigos de estado en el propio protocolo a analizar, la necesidad de conjuntos de datos iniciales válidos, y problemas con el análisis de protocolos que abren múltiples sockets.

Resultados, publicaciones y trabajo futuro

Durante este doctorado, se realizaron dos estancias de investigación (CEA-List en Francia y TU Delft en Países Bajos), se publicaron tres artículos en revistas Q1/Q3, otro artículo se encuentra en proceso de revisión, y se desarrollaron varias herramientas de código abierto. Durante esta tesis, también se realizaron aportes a diferentes proyecto estratégicos de ciberseguridad, se impartió docencia en asignaturas de grado y máster, se codirigió un TFG, y se contribuyó a la comunidad científica como revisor en varios congresos y revistas.

En cuanto a líneas futuras, se propone mejorar la cobertura de SYNOPTIC para manejar múltiples sockets y automatizar el análisis de ejecución simbólica con nuevos motores más escalables; e incluir mecanismos de reenvío en CHITON para tener mayor robustez frente a pérdida de paquetes UDP.

Índice de Autores

Abbaoui Toufiq, B., 398
Aguado-Caballero, I., 568, 570, 584
Aguirre, A., 348
Alaiz-Rodríguez, R., 563, 579
Alba Jiménez, C., 591
Alba Jiménez, C. M., 68
Alegre, E., 566, 568, 570, 572, 584
Alvarez, D., 340
Alvarez-Campana, M., 340
Amonarriz Pagola, I., 525
Anciaux, N., 42
Aparicio, A., 18, 26
Arrizabalaga, S., 597
Awad, E., 618

Bagheri-Gisour Marandyn, F., 604
Barreiro-Morante, A., 504
Barriga Rodríguez, Á., 59
Baños Ramos, A., 618
Beltrán-Lopez, P., 323
Betancourt Alonso, M., 105
Bifulco, R., 90
Blanco-Aza, D., 549
Blanco-Jiménez, F. J., 98
Blanco-Medina, P., 570
Blasco, J., 504
Blesa Gascón, A., 248
Bofill González, A., 381, 390
Bovet, G., 214
Braga, C. M., 529
Buitrago López, A., 130
Buitrago-Perez, M., 42
Bustamante V., D. F., 181

C. Hernández Goya, C., 490
Caballero, J., 436
Caballero-Gil, C., 276
Caballero-Gil, P., 255, 262, 269, 553
Cabuya-Padilla, D., 406
Calonge Fabregat, M., 442
Calvo, A., 560
Campo, C., 150
Caro Lindo, A., 381
Caro, A., 518
Carofilis, A., 570
Carrillo-Mondéjar, J., 141, 501, 600
Casares-Andrés, A., 98
Castaneda-Marroquín, C., 406, 481
Castelo Gómez, J. M., 600
Castillo-Fernández, E., 307
Castro Ponce, M., 604, 618
Cayuela Tudela, A. D., 130
Cillero, I., 542
Compastié, M., 560
Contreras, C., 513
Corchuelo, R., 364
Correa-Marichal, J., 553
Costin, A., 501
Criado-Lozano, P. A., 18, 26

Dafonte, C., 542
de Diego, S., 611
de Fuentes, J. M., 42
de la Cruz, A., 509
DeCastro-García, N., 73, 134, 283
del Rio-Álvarez, A. C., 504
Di Bartolo, A. J., 390
Diaz, 165
Domingo Martín, D., 544
Domínguez-Jiménez, J. J., 315
Díaz Bonete, R., 544
Díaz-López, D., 105, 222, 372, 406
Díaz-Redondo, R. P., 535
Díez-del-Río, E., 563, 566, 572

El Oualidi Charchmi, M., 134, 283
Escanez-Exposito, D., 255, 553
Escobar, S., 523
Escribano Ruíz, P., 372
Escrig, J., 560
Escuder Folch, S., 560
Escudero García, D., 73
Estepa Alonso, A., 165
Estepa Alonso, R., 165
Estepa, R., 473
Etxezarreta, X., 348

Feijóo-Añazco, A., 222
Fernández Becerra, L., 34
Fernández Carrasco, J. Á., 525

Fernández Isabel, A., 64
Fernández Llamas, M., 214
Fernández Maimó, L., 206
Fernández Rosales, C., 591
Fernández Rosales, C. , 68
Fernández Tárraga, M., 130
Fernández Vilas, A., 542
Fernández-Isabel, A., 513
Fernández-Jiménez, F. J., 473
Fernández-Medina, E., 181, 230, 331, 529
Fernández-Robles, L., 568, 572
Fernández-Veiga, M., 535, 542
Fidalgo, E., 563, 566, 568, 570, 572, 584
Figueroa, V., 331
Figueroa-Lorenzo, S., 597
Fontán Santamaría, A., 126
Franco de la Peña, M., 206
Frías-Guadalupe, S., 481
Fuentes Rodríguez, J., 126
Fúster, J., 90

Galán, D., 523
Gamiz, I., 82
Garcia De Alcaraz Ruiz, T., 130
Garcia-Campos, J. M., 473
Garcia-Diaz, J., 255
Garcia-Estevez, D., 622
Garcia-Ochoa, J., 64
García Cid, M. I., 544
García Clemente, F. J., 121
García Gil, S., 59
García Sánchez, S., 291
García Villalba, L. J., 449, 586
García, V., 523
García-Algora, D., 504
García-Arias, E., 9
García-Ochoa, J., 513
García-Pagán, C., 540
García-Rubio, C., 150
García-Sáez, L. M., 141
García-Teodoro, P., 98, 307
Garitano, I., 348
Gaul, C., 574, 584
Gañan, C. H., 602
Genua, E., 348
Gestal, M., 609
Gil Pérez, M., 323
Gomez, G., 436
Gonzalez, R., 90, 436
Gonzalez-Manzano, L., 42
González Fernández, I., 34
González Morales, V., 518
González Muñiz, A., 68
González Sánchez, B., 59
González-Castro, V., 574, 579
González-López, F., 98
González-Tablas, A. I., 414, 428
Gordillo Burrero, D., 381, 390
Guanciale, R., 299
Guarnieri, M., 299
Guerrero Hernández, J. M., 34
Guerrero-Higueras, Á. M., 50
Guillén-Mass, J. A., 198
Gutierrez-Aguero, I., 622
Gutiérrez Mlot, E. D., 602
Gutiérrez-Galeano, L., 315
Gómez López, M. T., 364
Gómez Mármol, F. , 121
Gómez-Goiri, A., 622
Gómez-Hernández, J. A., 307
Gómez-Pascual, A., 574

Hernantes, J., 597
Hernández Castro, J., 442
Hernández Martín, Á., 59
Hernández, R., 465
Hernández-Goya, C., 262
Homaei, M., 398, 518
Huertas Celdrán, A., 214, 457

Ibanez-Lissen, L., 42
Igual Catalán, R., 555
Inza Aldaz, J., 356
Iturbe, E., 527
Iturbe, M., 113, 348

Jacob, E., 82
Jiménez-Balsa, G., 540
Jover-Walsh, O., 291
Jáñez-Martino, F., 563, 568, 579

Kabir, F., 538
Koohpayeh Araghi, T., 538, 551
Kotsiuba, I., 602

Larriva-Novo, X., 291, 340
Llorente-Vazquez, O., 527
Loureiro-Acuña, J., 540
López Bernal, S., 457
López Cano, E., 64
López Espinosa, O., 64
López Gómez, A. M., 356
López López, G., 356, 604, 618
López Madejska, V. M., 457
López Martínez, A., 222

López Trigo, A., 381, 390
López, O., 513

Machaka, V., 597
Maciá-Fernández, G., 611
Magán-Carrión, R., 198
Mallen, L., 173
Marco Francia, M. P., 421, 613, 624
Marinaro, T., 299
Martinez, F., 157
Martinez, F. J., 173
Martín De Diego, I., 64
Martín de Diego, I., 513
Martínez Beltrán, E. T., 214
Martínez Hernández, L. A., 586
Martínez Pérez, G., 105, 130, 214, 222, 372, 457
Martínez, F., 230
Martínez, J. L., 141, 600
Martínez-González, M. M., 18, 26
Martínez-Luaña, X., 535, 540
Martínez-Mendoza, A., 563, 566, 568, 572, 574, 584
Marín-Caballero, M., 105
Matellán-Olivera, V., 50
Mateo, A., 340
Mayor, V., 473
Meadows, C., 523
Medina Merodio, J. A., 442
Megías, D., 538, 551
Meseguer, J., 523
Mogollón-Gutiérrez, Ó., 73, 518
Molina-Gil, J., 255, 276
Montero Ruiz, M. T., 421, 613
Moreno Sánchez, E., 299
Moreno, M., 113
Morillas, R., 68, 591
Muñoz Martín, S., 586
Muñoz Pérez, L. M., 555
Muñoz-Calle, J., 165, 473
Márquez Trujillo, A. G., 364
Méndez-Suárez, M., 9, 549

Nemati, H., 299
Nespoli, P., 105, 222, 323, 372
Nieto Acero, J., 59
Nóvoa, F. J., 542

Orduna Urrutia, R., 525
Ortega-Fernandez, I., 540
Ortiz, N., 560
Ortiz-Pradillo, J. C., 421, 624

P. Caballero-Gil, P., 490
Padilla Pío, J., 276
Padín-Torrente, H., 540
Paguada, S., 239
Parada Tallante, J., 631
Paredes, P., 481
Pastor Vargas, R., 465
Pastor-Galindo, J., 130
Pastor-Vargas, R., 9, 549
Pastrana Portillo, S., 511
Pastrana, S., 509
Patrignani, M., 299
Pazos, A., 609
Pegueroles, J., 189
Perales Gómez, Á. L., 105, 206, 222, 372
Perez Perez, D., 239
Perez, L., 340
Pilgermann, M., 189
Pinedo Terrados, G., 428
Pita Lorenzo, Y., 34
Povedano Álvarez, D., 449
Puigbó Sanvisens, J., 64
Pérez Arteaga, S., 586
Pérez Vázquez, B., 126
Pérez-Fuente, A., 18, 26
Pérez-Ramos, É., 262, 490

R. Ravines, R., 64
Rashed, M., 150
Ravines, R. R., 513
Rego, A., 527
Regueiro, C., 82, 611
Reina Quintero, A. M., 364
Reneses Botija, M., 618
Rios, E., 527
Rivera-Dourado, M., 609
Robles-Gómez, A., 9, 465, 549
Rodriguez Barredo-Valenzuela, A., 511
Rodriguez-Vega, M., 269
Rodríguez Fernández, R., 64
Rodríguez Lera, F. J., 34
Rodríguez Rodríguez, A. A., 356
Rodríguez, R. J., 501, 602
Rodríguez-Gómez, R. A., 198
Rodríguez-Lera, F. J., 50
Rodríguez-Pérez, N., 604
Roldán-Gómez, J., 141, 173, 600
Román Aguilar, R., 68
Román Martínez, J., 134, 283
Román, R., 525, 591
Rosado, D. G., 157, 181, 230, 331
Ruipérez Valiente, J. A. , 121
Ruipérez-Valiente, J. A., 130
Ruiz Catalán, J. L., 157, 230

Ruiz-Villafranca, S., 141, 600

Salamanca Estévez, I., 134, 283
Salas Heras, M., 414
Saldana, J., 602
Sancho Núñez, J. C., 73, 390
Sandoval Orozco, A. L., 421, 449, 586
Sanguesa, J. A., 173
Santos Anés, D., 59
Santos-Olmo, A., 157, 181, 230, 331
Sanvito, D., 90, 436
Sanz Fuertes, D., 248
Sequera Fernández, M. J., 398
Serrano, M. A., 529
Serón Arbeloa, F. J., 248
Sigrist, L., 604
Siracusano, G., 90, 436
Sobrín Hidalgo, D., 34
Sobrín-Hidalgo, D., 50
Soler, D., 542
Solera-Cotanilla, S., 291
Suarez-Tangil, G., 501, 511
Suárez-Bárcena, Á., 157
Syne, L., 262
Sáez-de-Cámara, X., 113
Sánchez Crespo, L. E., 157, 331
Sánchez González, F., 130
Sánchez, L. E., 181, 230
Sánchez-Corchado, Á., 315
Sánchez-Matas, A., 372
Sánchez-Zas, C., 291

Tanveer, W., 566
Tejada-Triviño, J. M., 307
Tejerina Rodríguez, O., 356
Tejerina, O., 42
Tobarra, L., 9, 465, 549
Toledo, N., 527
Torres-Sanz, V., 173
Triana Robles, D., 134, 283
Tuset-Peiró, P., 189

Urbieta, A., 113
Uroz Hinarejos, D., 641

Valcarce-Álvarez, R., 579
Vallina-Rodriguez, N., 511
van Liebergen, K., 436
Varela-Vaca, Á. J., 364
Vidal-Balboa, P., 9, 549
Vigoya, L., 68, 591
Vilajosana, X., 189
Villagrá, V. A., 291, 340

Vázquez-Naya, J., 609

Williams, R., 299

Yung, M., 276

Zine Boujemaoui, A., 214
Zurutuza, U., 348

Ávila Vegas, M. M., 398