

## Trabajo Fin de Grado

Clasificación automática de las etapas del sueño  
mediante técnicas de aprendizaje profundo  
Automatic sleep stage scoring using deep learning  
techniques

Autor

Emilio Estevan Tomás

Directora

María Sierra Torralba

Ponente

Luis Montesano del Campo

Grado en Ingeniería Informática

ESCUELA DE INGENIERÍA Y ARQUITECTURA  
2023



## Agradecimientos

---

Quiero expresar mi más sincero agradecimiento a todas las personas del equipo de Bitbrain que han formado parte de este trabajo, especialmente a mis supervisores Luis Montesano y María Sierra, por brindarme la excepcional oportunidad de participar en este proyecto de investigación y compartir conmigo su inmenso conocimiento. Su orientación profesional e incansable dedicación han sido fundamentales para el desarrollo del trabajo que tengo el orgullo de presentar en este documento.

También, quiero dar las gracias a mi familia, amigos y pareja por acompañarme y apoyarme incondicionalmente en los momentos más duros y frustrantes del proyecto y, en general, de mi carrera como estudiante. Nada de esto sería posible sin vosotros. Este trabajo también es vuestro.



## Resumen

---

El sueño juega un papel crucial en la salud mental y física de las personas. Dormir adecuadamente favorece el desarrollo de procesos neurocognitivos, como la consolidación de la memoria, el aprendizaje, la concentración y atención, el estado emocional y la toma de decisiones. Sin embargo, los trastornos del sueño representan un creciente y peligroso problema de salud pública en la sociedad en la que vivimos.

La polisomnografía (PSG) nocturna es el estándar de oro para la evaluación del sueño y se encarga de la medición en paralelo de múltiples señales fisiológicas. El primer paso esencial en el diagnóstico y tratamiento de los trastornos del sueño es la clasificación de las fases del sueño, que consiste en dividir los datos registrados por el PSG en segmentos de 30 segundos, denominados épocas, y etiquetarlos con uno de los estados del sueño. No obstante, se trata de un procedimiento muy costoso en tiempo que precisa de la intervención manual de técnicos especializados, traducándose en una tarea lenta, repetitiva y compleja que es invasiva en el sueño del paciente debido a la colocación de un elevado número de sensores en el cuerpo.

A partir de este contexto, el objetivo del presente trabajo consiste en el diseño y desarrollo de un modelo de aprendizaje automático para la clasificación de las etapas del sueño, automatizando la gestión de las señales recopiladas a través de dispositivos EEG vestibles encargados de la monitorización del sueño en entornos domésticos, evitando las dificultades y los costes existentes en la obtención de los datos con el PSG. Adicionalmente, debido al comportamiento de caja negra que presentan este tipo de algoritmos, se proponen técnicas de explicabilidad que posibiliten un mayor entendimiento acerca de su funcionamiento.

Para ello, partiendo del estudio de las metodologías que conforman el estado del arte en la medicina del sueño, se ha desarrollado una red neuronal profunda que ha sido evaluada en diferentes conjuntos de datos heterogéneos. Los resultados presentados en este proyecto demuestran la validez del modelo para clasificar el sueño en tiempo real a través de un número limitado de canales de información, cuyo rendimiento se encuentra a la altura del obtenido por evaluadores humanos profesionales. Además, se han validado los métodos de interpretabilidad propuestos, los cuales ayudan a la comprensión de cada uno de los componentes de la arquitectura del modelo y, en consecuencia, facilitan la aceptación y adopción de esta tecnología en el sector clínico. De esta forma, el presente trabajo sirve como prueba de concepto para una futura tecnología del sueño que permitirá reducir la carga de trabajo que sufren muchos médicos y escalar el estudio del sueño a las necesidades de millones de personas en el mundo que padecen trastornos del sueño.



# Índice

|          |                                                            |           |
|----------|------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introducción</b>                                        | <b>1</b>  |
| 1.1      | Arquitectura del sueño . . . . .                           | 3         |
| 1.2      | Trastornos del sueño . . . . .                             | 4         |
| 1.3      | Electroencefalografía (EEG) . . . . .                      | 5         |
| 1.3.1    | Formas de onda . . . . .                                   | 5         |
| 1.3.2    | Aspectos técnicos . . . . .                                | 5         |
| 1.4      | Propiedades electroencefalográficas del sueño . . . . .    | 7         |
| 1.5      | Objetivos y alcance del proyecto . . . . .                 | 9         |
| <b>2</b> | <b>Trabajo relacionado</b>                                 | <b>10</b> |
| 2.1      | Clasificación automática de las etapas del sueño . . . . . | 10        |
| 2.1.1    | Shallow learning . . . . .                                 | 10        |
| 2.1.2    | Deep learning . . . . .                                    | 11        |
| 2.2      | Explicabilidad e interpretabilidad . . . . .               | 14        |
| <b>3</b> | <b>Metodología</b>                                         | <b>16</b> |
| 3.1      | Arquitectura del modelo . . . . .                          | 16        |
| 3.2      | Técnicas de explicabilidad e interpretabilidad . . . . .   | 19        |
| 3.2.1    | Cuantificación de la confianza . . . . .                   | 19        |
| 3.2.2    | Análisis probabilístico . . . . .                          | 20        |
| 3.2.3    | Influencia de las épocas vecinas . . . . .                 | 20        |
| 3.2.4    | Grad-CAM . . . . .                                         | 21        |
| 3.3      | Evaluación . . . . .                                       | 22        |
| 3.3.1    | Datos disponibles . . . . .                                | 22        |
| 3.3.2    | Rendimiento . . . . .                                      | 24        |
| <b>4</b> | <b>Resultados y discusión</b>                              | <b>28</b> |
| 4.1      | Clasificación automática de las fases del sueño . . . . .  | 28        |
| 4.1.1    | Datasets públicos . . . . .                                | 28        |
| 4.1.2    | Datos recopilados por Bitbrain . . . . .                   | 33        |
| 4.2      | Explicabilidad e interpretabilidad . . . . .               | 35        |
| <b>5</b> | <b>Conclusiones</b>                                        | <b>42</b> |

|                                                                          |           |
|--------------------------------------------------------------------------|-----------|
| 5.1 Trabajo futuro . . . . .                                             | 43        |
| <b>Bibliografía</b>                                                      | <b>51</b> |
| <b>Apéndices</b>                                                         | <b>52</b> |
| <b>A Gestión del proyecto</b>                                            | <b>53</b> |
| A.1 Herramientas software y hardware . . . . .                           | 53        |
| A.2 Planificación . . . . .                                              | 53        |
| <b>B Resultados de los métodos de codificación posicional</b>            | <b>54</b> |
| <b>C Estudio de hiperparámetros</b>                                      | <b>55</b> |
| <b>D Resultados con el modelo SleepXAI</b>                               | <b>57</b> |
| <b>E Resultados sobre el dataset ISRUC-Sleep</b>                         | <b>60</b> |
| <b>F Resultados adicionales sobre explicabilidad e interpretabilidad</b> | <b>62</b> |

# Lista de Figuras

|     |                                                                                                                                                                                                  |    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | Ilustración de las configuraciones presentadas para la monitorización del sueño.                                                                                                                 | 2  |
| 1.2 | Hipnograma asociado a la distribución y evolución típica de las etapas sueño a lo largo de la noche. . . . .                                                                                     | 4  |
| 1.3 | Clasificación de las formas de onda del EEG según su frecuencia. . . . .                                                                                                                         | 5  |
| 1.4 | Posiciones de los electrodos EEG en el sistema 10-10, junto con los fiduciales y los lóbulos cerebrales asociados. . . . .                                                                       | 6  |
| 1.5 | Propiedades del EEG en el estado de vigilia. . . . .                                                                                                                                             | 7  |
| 1.6 | Propiedades del EEG en la fase N1 del sueño. . . . .                                                                                                                                             | 7  |
| 1.7 | Propiedades del EEG en la fase N2 del sueño. . . . .                                                                                                                                             | 8  |
| 1.8 | Propiedades del EEG en la fase N3 del sueño. . . . .                                                                                                                                             | 8  |
| 1.9 | Propiedades del EEG en la fase REM del sueño. . . . .                                                                                                                                            | 8  |
| 2.1 | Esquema secuencia a secuencia ( <i>sequence-to-sequence</i> ) utilizado en redes neuronales para la clasificación automática del sueño [24]. . . . .                                             | 12 |
| 3.1 | Arquitectura del modelo de aprendizaje profundo. . . . .                                                                                                                                         | 17 |
| 3.2 | Datasets disponibles para la evaluación del modelo de clasificación del sueño.                                                                                                                   | 23 |
| 3.3 | Esquema de los pasos para la evaluación del método propuesto. . . . .                                                                                                                            | 24 |
| 4.1 | Matrices de confusión para cada test realizado sobre el dataset SleepEDF. . .                                                                                                                    | 29 |
| 4.2 | Matrices de confusión para cada test realizado sobre el dataset DOD-H. . . .                                                                                                                     | 30 |
| 4.3 | Matrices de confusión para cada test realizado sobre el dataset DOD-O. . . .                                                                                                                     | 32 |
| 4.4 | Matrices de confusión para cada test realizado sobre el dataset Study 1 de Bitbrain. . . . .                                                                                                     | 34 |
| 4.5 | Matrices de confusión para cada test realizado sobre el dataset Study 3 de Bitbrain. . . . .                                                                                                     | 35 |
| 4.6 | Ilustración de la estimación de la confianza del modelo en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG ( <i>accuracy</i> = 88.35%, evaluación <i>offline</i> ). . . . . | 36 |
| 4.7 | Análisis de las probabilidades de salida del modelo en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG ( <i>accuracy</i> = 88.35%, evaluación <i>offline</i> ). . . . .     | 37 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.8 | Resultados obtenidos mediante la aplicación del método Grad-CAM sobre el modelo (segunda capa convolucional de la red), indicando las regiones de las épocas que más activación experimentan en relación con las reglas definidas en el manual de la AASM, correspondientes a cada fase del sueño. La ejecución se ha realizado en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG ( <i>accuracy</i> = 88.35%, evaluación <i>offline</i> ). . . . . | 39 |
| 4.9 | Resultados del análisis de la influencia de las épocas vecinas respecto a la clasificación de la época objetivo (segunda capa Multi-Head Self-Attention de la red) en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG ( <i>accuracy</i> = 88.35%, evaluación <i>offline</i> ). . . . .                                                                                                                                                              | 41 |
| A.1 | Diagrama de Gantt del proyecto desarrollado. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                     | 53 |
| D.1 | Arquitectura del modelo SleepXAI. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                | 57 |
| D.2 | Matrices de confusión para cada test realizado con el modelo SleepXAI. . . .                                                                                                                                                                                                                                                                                                                                                                                             | 59 |
| E.1 | Matrices de confusión para cada test realizado sobre el dataset ISRUC-Sleep. .                                                                                                                                                                                                                                                                                                                                                                                           | 61 |
| F.1 | Ilustración de la estimación de la confianza del modelo en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( <i>accuracy</i> = 87.08%, evaluación <i>offline</i> ). . . . .                                                                                                                                                                                                                                                                           | 63 |
| F.2 | Análisis de las probabilidades de salida del modelo en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( <i>accuracy</i> = 87.08%, evaluación <i>offline</i> ). . . . .                                                                                                                                                                                                                                                                               | 64 |
| F.3 | Resultados obtenidos mediante la aplicación del método Grad-CAM sobre el modelo (segunda capa convolucional de la red), indicando las regiones de las épocas que más activación experimentan en relación con las reglas definidas en el manual de la AASM, correspondientes a cada fase del sueño. La ejecución se ha realizado en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( <i>accuracy</i> = 87.08%, evaluación <i>offline</i> ). . . . .   | 65 |
| F.4 | Resultados del análisis de la influencia de las épocas vecinas respecto a la clasificación de la época objetivo (primera capa Multi-Head Self-Attention de la red) en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( <i>accuracy</i> = 87.08%, evaluación <i>offline</i> ). . . . .                                                                                                                                                                | 66 |

# Lista de Tablas

|     |                                                                                                                                                                                                                                   |    |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | Etapas del sueño y duración asociada. . . . .                                                                                                                                                                                     | 3  |
| 2.1 | Recopilación de modelos de <i>deep learning</i> para la clasificación automática de las etapas del sueño. . . . .                                                                                                                 | 13 |
| 3.1 | Resumen de las pruebas planteadas sobre los datasets de Bitbrain. . . . .                                                                                                                                                         | 26 |
| 3.2 | Resumen de las pruebas planteadas sobre los datasets públicos. . . . .                                                                                                                                                            | 27 |
| 4.1 | Resultados del informe de clasificación sobre el dataset SleepEDF. . . . .                                                                                                                                                        | 29 |
| 4.2 | Resultados del informe de clasificación sobre el dataset DOD-H. . . . .                                                                                                                                                           | 30 |
| 4.3 | Resultados del informe de clasificación sobre el dataset DOD-O. . . . .                                                                                                                                                           | 31 |
| 4.4 | Resultados del informe de clasificación sobre el dataset Study 1 de Bitbrain. .                                                                                                                                                   | 33 |
| 4.5 | Resultados del informe de clasificación sobre el dataset Study 3 de Bitbrain. .                                                                                                                                                   | 34 |
| 4.6 | Resumen de las reglas visuales del manual de la AASM. . . . .                                                                                                                                                                     | 38 |
| B.1 | Resultados de cada método de codificación posicional sobre el dataset SleepEDF utilizando la información del canal Fpz-Cz. . . . .                                                                                                | 54 |
| C.1 | Estudio de hiperparámetros sobre el dataset <i>Study 3</i> de Bitbrain en los canales AF7 + AF8 de la banda textil. La primera fila se corresponde con el modelo base y las celdas vacías con los parámetros por defecto. . . . . | 56 |
| C.2 | Estudio de hiperparámetros sobre el dataset <i>Study 3</i> de Bitbrain en los canales F3 + F4 del PSG. La primera fila se corresponde con el modelo base y las celdas vacías con los parámetros por defecto. . . . .              | 56 |
| C.3 | Estudio de hiperparámetros sobre el dataset <i>Study 1</i> de Bitbrain en los canales F3 + F4 del PSG. La primera fila se corresponde con el modelo base y las celdas vacías con los parámetros por defecto. . . . .              | 56 |
| D.1 | Resultados de la ejecución de las pruebas sobre el modelo SleepXAI. . . . .                                                                                                                                                       | 58 |
| E.1 | Resultados del informe de clasificación sobre el dataset ISRUC-Sleep. . . . .                                                                                                                                                     | 60 |

# Capítulo 1

## Introducción

El sueño constituye prácticamente un tercio de nuestras vidas y resulta fundamental en la salud mental y física de las personas [1], [2]. La calidad del sueño influye en múltiples funciones fisiológicas del ser humano, incluyendo los sistemas inmunitario, metabólico y cardiovascular [3]-[5]. Dentro del complejo neurobiológico que es el cerebro, dormir adecuadamente favorece el desarrollo de procesos neurocognitivos, como la consolidación de la memoria, el aprendizaje, la concentración y atención, el estado emocional y la toma de decisiones [6]-[9]. La privación del sueño aumenta el riesgo de sufrir enfermedades cardiovasculares y metabólicas, la mortalidad y la aparición de patologías relacionadas con el Alzheimer [4], [10]-[12]. Sin embargo, los trastornos del sueño representan un creciente y peligroso problema de salud pública [13]. Por ejemplo, Ohayon y Smirne determinaron que un 27.6% de los ciudadanos italianos presentaba síntomas relacionados con alteraciones del sueño [14]. No obstante, menos del 20% de los casos son diagnosticados y tratados [15].

La polisomnografía (PSG) nocturna es el estándar de oro para la evaluación del sueño y, en consecuencia, la identificación de los trastornos del sueño [16]-[19]. Se encarga de la medición en paralelo de una amplia gama de señales fisiológicas, como el electroencefalograma (EEG), electrocardiograma (ECG), electrooculograma (EOG), electromiograma (EMG), la oxigenación sanguínea, el flujo de aire y el esfuerzo respiratorio. Estos datos son mayoritariamente recopilados en un entorno médico controlado, donde los pacientes permanecen durante toda la noche.

Uno de los primeros pasos esenciales para el estudio del sueño es identificar las fases del sueño experimentadas por el sujeto durante la noche, lo cual se conoce como el proceso de clasificación del sueño, en el que los datos registrados por el PSG son divididos en segmentos de 30 segundos, denominados *épocas*, y etiquetados con una de las etapas del sueño, obteniendo el espectro completo de las fases presentes en una grabación. Se trata de un procedimiento muy costoso en tiempo que precisa la intervención manual de técnicos especializados para realizar el análisis visual de las series temporales recopiladas, requiriendo de hasta 2 horas para analizar un PSG de 8 horas [20], traducándose en una tarea repetitiva, lenta y compleja. Además, los resultados de este proceso están sujetos a fallos subjetivos, ya que el grado de acuerdo entre evaluadores está entre el 70% y 80% [21], [22], y el mismo individuo presenta una consistencia del 90% para la misma grabación [23]. También, la información de una sola noche puede no ser representativa en cuanto a la patología que realmente presenta el sujeto analizado [24]. Esta técnica es invasiva en el sueño del paciente, ya que consiste en una prueba multiparamétrica que requiere la colocación de un elevado número de sensores en el cuerpo del sujeto, y presenta un coste temporal y económico tan elevado que impide una escalabilidad que responda a las necesidades de millones de personas que sufren trastornos del sueño.

Por lo tanto, teniendo en cuenta la relevancia vital del sueño en la salud humana y las dificultades existentes en el procedimiento de evaluación del sueño, surge la necesidad de simplificar y reducir el coste de la técnica descrita. En consecuencia, Bitbrain Technologies, una empresa de neurotecnología que nació en 2010 como spin-off de la Universidad de Zaragoza y que combina neurociencia e inteligencia artificial, está desarrollando un nuevo dispositivo hardware ergonómico y portátil, compuesto de materiales textiles inteligentes para monitorizar el sueño en entornos domésticos mediante la medición del electroencefalograma (EEG) a través de 5 canales [25], siendo posible obtener una precisión médica utilizando únicamente el EEG [26], [27].

De esta manera, se evitan los problemas expuestos, en relación con las dificultades existentes en la obtención de los datos con el PSG, sin la necesidad del entorno médico al que se somete el paciente ni los consecuentes costes derivados. La finalidad del sistema no reside únicamente en mejorar el diagnóstico de los trastornos del sueño, sino también en realizar intervenciones en bucle cerrado (*Closed-Loop Stimulation*) que cubren diversas aplicaciones, como el aumento y mejora de las capacidades cognitivas [28], [29]. En la Figura 1.1 se puede observar la configuración de las alternativas presentadas.

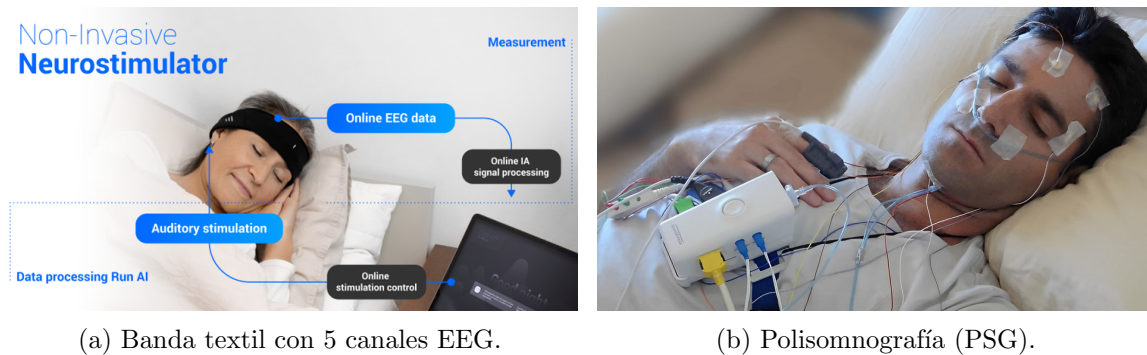


Figura 1.1: Ilustración de las configuraciones presentadas para la monitorización del sueño.

Esta tecnología debe ser simple, barata y accesible para el mayor número de personas, puesto que está diseñada para operar en espacios domésticos por usuarios no especializados en la materia. Por consiguiente, la calidad de la señal recopilada se verá influenciada por la presencia de artefactos y ruido, debido a que la complejidad del dispositivo es menor respecto a la medición obtenida con el PSG, en cuanto al número de sensores y a la ausencia de geles conductores, lo cual supone un reto adicional.

En definitiva, a partir de este contexto y motivación, y en colaboración con Bitbrain Technologies, el presente trabajo pretende proponer una técnica de automatización de la gestión de los datos recopilados para llevar a cabo la clasificación del sueño mediante el uso de técnicas aprendizaje profundo (*deep learning*). Adicionalmente, dado que una de las aplicaciones consiste en llevar a cabo intervenciones en bucle cerrado, este algoritmo también debe ser capaz de evaluar el sueño en tiempo real. Para ello, se llevará a cabo un estudio de las metodologías presentes en el estado del arte, de las cuales se desarrollarán e implementarán las que se consideren más adecuadas, que serán evaluadas sobre diferentes conjuntos de datos para determinar su rendimiento y efectividad, así como el grado de cumplimiento en el procedimiento de evaluación del sueño respecto a la aplicación médica resultante.

Sumado a todo esto, debido a que este tipo de algoritmos presentan un comportamiento de caja negra, resulta fundamental en el ámbito sanitario demostrar la fiabilidad de las técnicas empleadas, principalmente en cuanto al razonamiento con el que los modelos toman las decisiones, de forma que también se aplicarán técnicas de interpretabilidad y explicabilidad que posibiliten un mayor entendimiento de su funcionamiento para facilitar la aceptación y adopción de esta tecnología en el sector médico [30].

Los aspectos descritos se detallarán en los próximos capítulos del informe. Previamente, se van a establecer los fundamentos teóricos necesarios para el entendimiento del trabajo, en cuanto a la arquitectura del sueño, los principales trastornos asociados, la electroencefalografía (EEG) y las propiedades electroencefalográficas del sueño, en las secciones 1.1, 1.2, 1.3 y 1.4, respectivamente.

## 1.1. Arquitectura del sueño

El sueño constituye la suspensión periódica natural de la conciencia, caracterizada por una inmovilidad relativa y una capacidad de respuesta reducida a estímulos externos [31]. Normalmente, este periodo tiene una duración de entre 7 y 9 horas, en el que existen dos grandes fases que se van alternando durante la noche: NREM (*Non-rapid Eye Movement sleep*) y REM (*Rapid Eye Movement sleep*).

La primera parte de la noche está dominada por el sueño de ondas lentas (*slow-wave sleep*, SWS), mientras que el sueño REM prevalece durante la segunda mitad [6]. La fase NREM se compone, a su vez, de 4 etapas distintas (*stage I*, *stage II*, *stage III - IV*) siguiendo las reglas del manual de Rechtschaffen and Kales [32], establecido en 1968. Sin embargo, en 2007 la Academia Americana de Medicina del Sueño (*American Academy of Sleep Medicine*, AASM) definió que la fase NREM se compone de 3 estados (*N1*, *N2*, *N3*) [33], resultando en la convención utilizada en los trabajos de clasificación automática del sueño [34] y, en consecuencia, la que ha sido adoptada en este proyecto. En la Tabla 1.1 se recogen las diferentes etapas con la duración asociada.

| Fase del sueño             | Tipo | Duración      | Total en la noche |
|----------------------------|------|---------------|-------------------|
| <b>N1</b> (Stage I)        | NREM | 5-10 minutos  | 2-5%              |
| <b>N2</b> (Stage II)       | NREM | 20-25 minutos | 50%               |
| <b>N3</b> (Stage III - IV) | NREM | 20-40 minutos | 20-50%            |
| <b>REM</b> (Stage V)       | REM  | 10-60 minutos | 20-25%            |

Tabla 1.1: Etapas del sueño y duración asociada.

La clasificación del sueño se realiza asignando una de las etapas a cada época (segmento de 30 segundos) que comprende el estudio del sueño del paciente. La representación gráfica de la distribución y evolución de las fases del sueño a lo largo de la noche se denomina *hipnograma* (ver Figura 1.2).



### 1.3. Electroencefalografía (EEG)

La electroencefalografía (EEG) constituye la medición no invasiva de los campos eléctricos del cerebro a través de electrodos colocados en el cuero cabelludo, los cuales registran potenciales postsinápticos de voltaje consecuentes del flujo de corriente presente en las neuronas [42]. Estas diferencias de potencial, computadas en  $\mu V$ , se calculan utilizando dos electrodos: uno para la monitorización y otro de referencia. La técnica fue realizada por primera vez en 1929 [43], y desde entonces ha abarcado un amplio espectro de aplicaciones, constituyendo el sensor principal del PSG para determinar los estados del sueño. Adicionalmente, en la sección 1.3.1 se explican las principales formas de onda presentes en estas señales, y en la sección 1.3.2 se abarcan los aspectos básicos que afectan a la configuración de un dispositivo EEG.

#### 1.3.1. Formas de onda

Las ondas cerebrales recopiladas a través del EEG se miden en hercios (Hz), o ciclos por segundo, y se clasifican según su frecuencia. En la literatura se distinguen 4 principales bandas de frecuencia: delta, theta, alpha y beta. Las ondas delta (1-4 Hz) predominan durante el sueño profundo. Las ondas theta (4-8 Hz) están implicadas en las funciones de memoria, mientras que las ondas alpha (8-12 Hz) intervienen en diferentes funciones motoras y cognitivas. Por último, las ondas beta (12-30 Hz) representan un indicador de la transmisión cortical inhibitoria, asociado a una concentración activa. En la Figura 1.3 se ilustran las diferentes formas de onda que caracterizan el EEG.

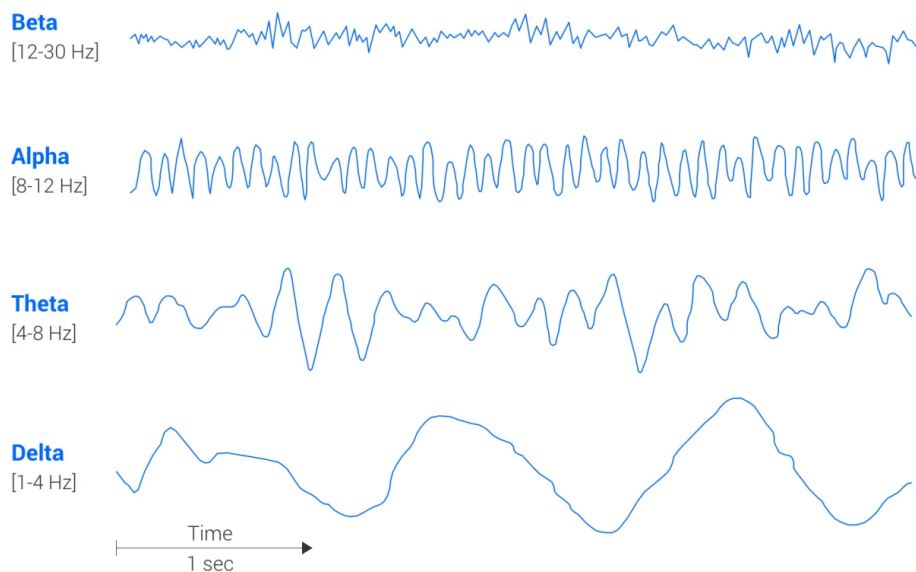


Figura 1.3: Clasificación de las formas de onda del EEG según su frecuencia.

#### 1.3.2. Aspectos técnicos

La configuración de un equipo EEG comprende la selección de los parámetros que constituyen cada uno de sus componentes. El primero de estos elementos es el amplificador, en el que es posible modificar aspectos como la frecuencia de muestreo, la resolución o el rango de entrada. Este dispositivo realiza el aumento y conversión de las señales eléctricas analógicas, captadas por los sensores, en señales digitales que pueden ser procesadas por un computador.

Por otro lado, se encuentra la capa de sensores, la cual contiene los electrodos que se colocan en el cuero cabelludo, encargados de capturar las señales eléctricas generadas por la actividad cerebral, de la que, dada su importancia, cabe destacar los siguientes parámetros:

- **Número de electrodos:** determina la cantidad de información que es posible medir y, en consecuencia, el nivel de detalle obtenido de las diferentes zonas del cerebro. Normalmente, este valor comprende el rango entre 8 y 128, refiriéndose al número de electrodos de registro de señal, sin tener en cuenta el electrodo de referencia (necesario para restar el ruido común a los electrodos de registro) y el electrodo de tierra (fundamental en la reducción de las interferencias y la garantía de la seguridad del paciente).
- **Colocación de los sensores:** establece los puntos del cuero cabelludo en los que se colocan los electrodos, comprendiendo las diferentes áreas del cráneo: prefrontal (Fp), frontal (F), temporal (T), parietal (P), occipital (O) y central (C). También, existe una zona (Z), situada en el plano sagital de la línea media del cráneo, así como un área (A), a veces denominada (M) por el mastoideo, ubicada en la apófisis ósea.

Una de las distribuciones más utilizadas es el sistema 10-20, en el que la distancia entre electrodos adyacentes es el 10% o el 20% de la distancia total anterior-frontal o izquierda-derecha del cráneo, de manera que se garantiza una localización uniforme y estándar en todos los sujetos, independientemente del tamaño de su cabeza. Esta métrica limita el número de electrodos que es posible colocar, de forma que existen otras posibles combinaciones, como el sistema 10-10 o el sistema 10-5, que establecen un espacio entre electrodos menor, permitiendo aumentar la cantidad de sensores colocados, así como la precisión del dispositivo. En la Figura 1.4 se puede observar la configuración de los electrodos para el sistema 10-10, así como los fiduciales y los lóbulos cerebrales implicados.

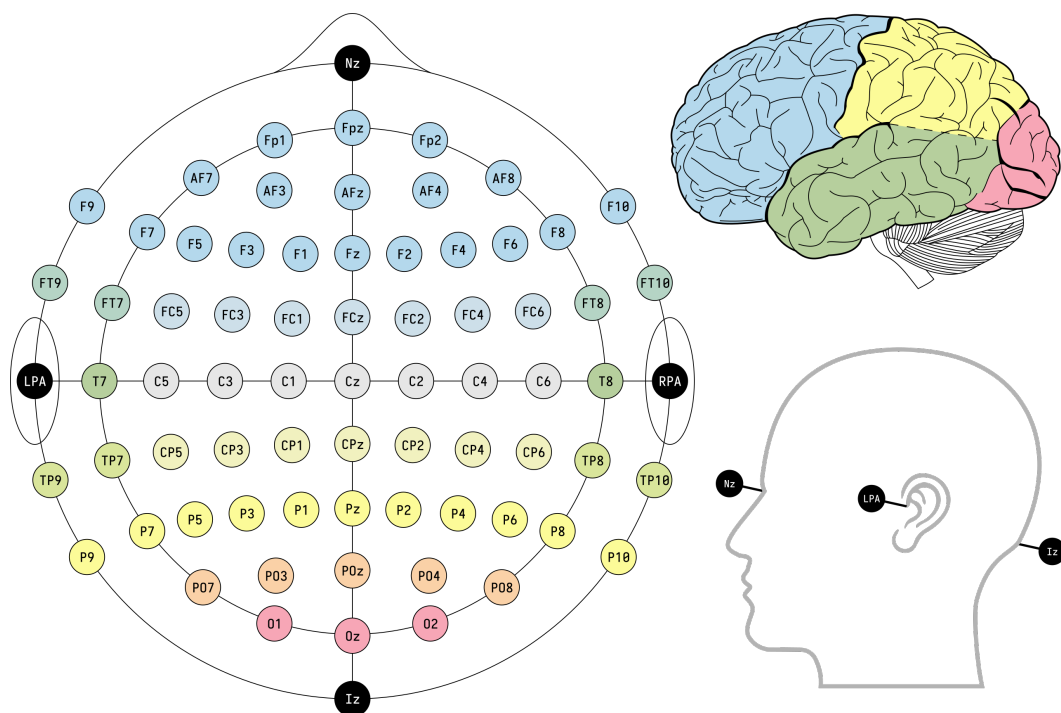


Figura 1.4: Posiciones de los electrodos EEG en el sistema 10-10, junto con los fiduciales y los lóbulos cerebrales asociados.

- **Contacto entre los electrodos y el cuero cabelludo:** existen dos grandes grupos según el medio de contacto entre los electrodos y el cráneo. En primer lugar, los electrodos húmedos requieren de una sustancia electrolítica (geles, soluciones salinas o agua) para mejorar la impedancia de contacto. Por otro lado los electrodos secos no precisan de estas sustancias, estableciendo un contacto directo, que suele implicar el uso de amplificadores con especificaciones técnicas superiores respecto a los húmedos para conseguir una calidad de señal similar a cambio de una configuración más simple.

## 1.4. Propiedades electroencefalográficas del sueño

Los expertos realizan la clasificación de las etapas del sueño siguiendo las reglas definidas por la AASM [33], la cual establece los patrones de onda y características de las frecuencias del EEG (sección 1.3.1) experimentadas en cada fase del sueño, que pueden ser descritas de la siguiente forma:

- **Estado de vigilia (Wake):** se caracteriza por la presencia de más del 50% de frecuencias alpha, generalmente sobre la región occipital, junto con eventos de parpadeo de los ojos y movimientos oculares rápidos o de lectura [34] (ver Figura 1.5).

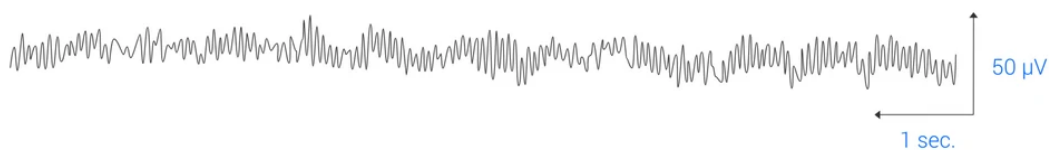


Figura 1.5: Propiedades del EEG en el estado de vigilia.

- **Fase N1:** muestra bandas de frecuencia alpha que no deben superar el 50% del espectro total, con la aparición de ondas theta de amplitud baja y movimientos oculares lentos [44] (ver Figura 1.6).

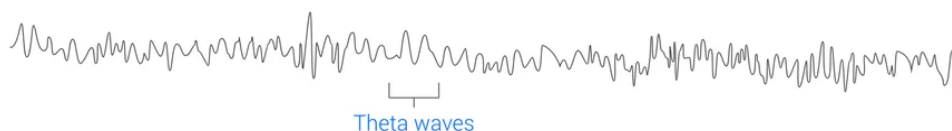


Figura 1.6: Propiedades del EEG en la fase N1 del sueño.

- **Fase N2:** exhibe los denominados husos del sueño (*sleep spindles*, 11-15 Hz), compuestos de ondas crecientes y menguantes que duran aproximadamente 0,5 segundos [31], en conjunto con eventos complejos K (*K-complex*), formados por deflexiones negativas (*downstate*) seguidas de una deflexión positiva menos intensa (*upstate*), con una duración aproximada de 1 segundo [45] (ver Figura 1.7).

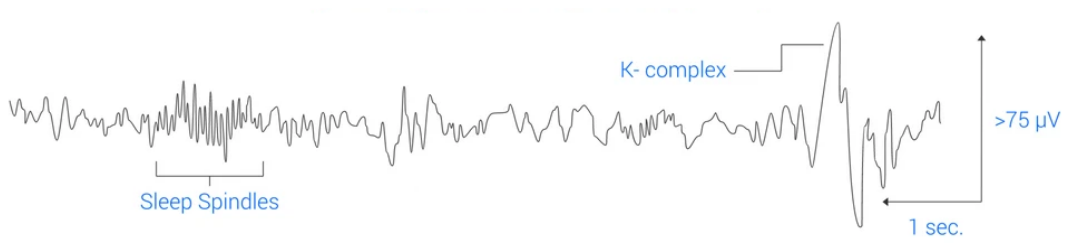


Figura 1.7: Propiedades del EEG en la fase N2 del sueño.

- **Fase N3:** predominan las ondas delta en más del 20% de la época [46], con la posible aparición de husos del sueño y complejos K [47] (ver Figura 1.8).

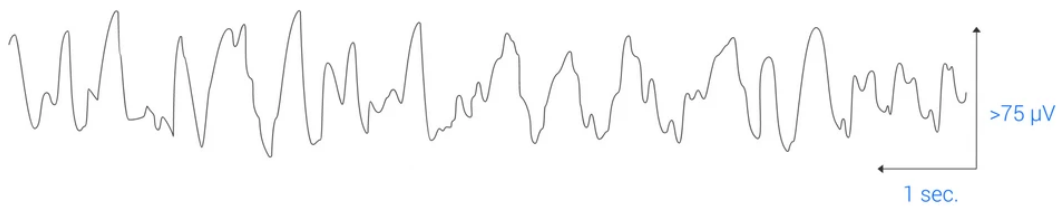


Figura 1.8: Propiedades del EEG en la fase N3 del sueño.

- **Fase REM:** manifiesta una actividad rápida y de bajo voltaje, compuesta de un aplanamiento del EEG en el rango de frecuencias alfa y theta [48]. Además, revela un tipo concreto de ondas theta, conocidas como ondas en diente de sierra (*sawtooth waves*), debido a su parecido con la hoja de una sierra, las cuales preceden al inicio de los movimientos oculares rápidos que dan nombre a esta etapa [49] (ver Figura 1.9).



Figura 1.9: Propiedades del EEG en la fase REM del sueño.

## 1.5. Objetivos y alcance del proyecto

El principal objetivo del presente trabajo consiste en el diseño y desarrollo de un modelo de aprendizaje automático para la clasificación de las etapas del sueño, automatizando la gestión de las señales recopiladas a través de dispositivos EEG vestibles encargados de la monitorización del sueño en entornos domésticos. Adicionalmente, debido al comportamiento de caja negra que presentan este tipo de algoritmos, se aplicarán técnicas de explicabilidad e interpretabilidad que posibiliten un mayor entendimiento acerca de su funcionamiento, con el propósito de facilitar la aceptación y adopción de esta tecnología en el sector clínico. Por consiguiente, se plantean los siguientes objetivos específicos:

- Análisis del conjunto de técnicas de aprendizaje automático existentes en el estado del arte para la evaluación automática de las fases sueño, junto con el estudio de los métodos de explicabilidad e interpretabilidad presentes en la literatura en relación con la medicina del sueño (capítulo 2).
- Diseño e implementación de un modelo de aprendizaje automático robusto y eficiente para la clasificación de las etapas del sueño (capítulo 3, sección 3.1).
- Implementación de técnicas de explicabilidad e interpretabilidad aplicadas sobre la arquitectura del modelo desarrollado (capítulo 3, sección 3.2).
- Evaluación del rendimiento del modelo sobre diferentes conjuntos de datos heterogéneos e independientes, entre los cuales destaca un dataset recopilado a través de la banda textil de Bitbrain, con el fin de experimentar con las metodologías de evaluación *offline* y en tiempo real, así como con la posición y el número de electrodos utilizados como fuentes de información (capítulo 3, sección 3.3 y capítulo 4).
- Análisis de los resultados obtenidos a través de la ejecución del modelo y validación de los métodos de explicabilidad e interpretabilidad implementados (capítulo 4).

## Capítulo 2

# Trabajo relacionado

La inteligencia artificial (IA) consiste en la simulación avanzada de los mecanismos cognitivos humanos, realizada a través de sistemas computacionales y algoritmos que son capaces de completar tareas que precisan de inteligencia humana. El aprendizaje automático (*machine learning*, ML) comprende una aplicación de la IA en la que los sistemas son capaces de aprender y mejorar automáticamente a partir de la experiencia, exhibiendo la habilidad de replicar procesos llevados a cabo por seres humanos para facilitar y simplificar determinados procedimientos manuales [34].

Este enfoque puede aplicarse en la clasificación de las etapas del sueño, que comprende un trabajo tedioso y repetitivo para los profesionales. Por consiguiente, el objetivo del presente capítulo consiste en describir el conjunto de técnicas de aprendizaje automático utilizadas para la automatización de la evaluación del sueño, correspondiente al apartado 2.1, así como las principales metodologías sobre explicabilidad e interpretabilidad existentes en este tipo de algoritmos, presentadas en la sección 2.2.

### 2.1. Clasificación automática de las etapas del sueño

Existen dos vertientes principales para abordar el problema de la clasificación automática de las etapas del sueño. En primer lugar, en el apartado 2.1.1 se exhibe el denominado aprendizaje superficial (*shallow learning*), basado en la extracción de características a partir del conocimiento de expertos. Por otro lado, en el apartado 2.1.2 se establecen los principales modelos de aprendizaje profundo (*deep learning*) existentes en la medicina del sueño, los cuales emplean directamente los datos en bruto durante el proceso de aprendizaje.

#### 2.1.1. Shallow learning

El aprendizaje superficial o *shallow learning* requiere la extracción manual de características (*features*) y la selección de clasificadores. El procedimiento consiste, en primer lugar, en el preprocesado y filtrado de los datos para obtener señales limpias. Después, se extraen las características y se seleccionan aquellas que se consideran relevantes para el modelo. Finalmente, se introducen las características elegidas en el clasificador para ejecutar la evaluación de las fases del sueño [50].

El espectro de las características extraídas de los datos incluye características del dominio del tiempo [51], características del dominio de la frecuencia [52] y características del dominio tiempo-frecuencia [53]. En cuanto a los métodos de clasificación, destaca el uso de los siguientes algoritmos y técnicas: K-medias (*K-means*) [54], Random Forest (RF) [55], Support

Vector Machine (SVM) [56], Bootstrap Aggregating [57], K-Nearest Neighbors (KNN) [58] y AdaBoost [59].

Concretamente, Guo et al. [60] desarrollaron un clasificador SVM fundamentado en la transformada de Hilbert-Huang y las características de entropía de muestra, mientras que en [61] los autores proponen un método basado en el análisis de componentes principales multiescala y la transformada Wavelet discreta, todo ello integrado en un clasificador RotSVM. Asimismo, Timplalexis et al. [62] obtuvieron características mixtas relacionadas con el dominio del tiempo y el dominio de la frecuencia, conformando un clasificador que opera mediante un mecanismo de votación.

Sin embargo, los métodos descritos tienen un rendimiento limitado, ya que han sido evaluados sobre conjuntos de datos de tamaño reducido, de forma que no se garantiza su efectividad en un contexto que cuenta con una cantidad de información mayor y en el que los sujetos son heterogéneos. Además, debido a que la extracción de características de los datos parte del conocimiento de expertos, se constituye como una metodología que solo puede ser llevada a cabo por profesionales formados en la materia. Por ello, la aplicación de técnicas de aprendizaje profundo (apartado 2.1.2) ha producido un impacto significativo en la evaluación del sueño, ya que supera muchas de las limitaciones que presentan los enfoques anteriores, arrojando resultados considerablemente superiores.

### 2.1.2. Deep learning

El aprendizaje profundo o *deep learning* se basa en el uso de redes neuronales artificiales (*artificial neural networks*, ANNs), diseñadas para imitar la forma en la que los cerebros biológicos resuelven los problemas, las cuales cuentan con múltiples capas de nodos de procesamiento de información, con un nivel de abstracción y complejidad incremental, a través de las que los datos fluyen. Una red neuronal se define a través de su arquitectura, que comprende aspectos como el número y tipo de capas, el número de neuronas, el tipo de conexión entre capas o las funciones de activación y pérdida, entre otros. La posibilidad de aprender óptimamente a partir de grandes conjuntos de datos conforma una de las principales razones para su aplicación en la clasificación del sueño.

Las primeras aproximaciones se situan entorno al año 2017, en las que se se utilizaba como entrada una ventana de épocas de tamaño reducido, ubicadas alrededor de la época del sueño a predecir [24]. De esta forma, se propusieron diferentes arquitecturas de redes autónomas, como redes neuronales profundas (*Deep Neural Networks*, DNNs) [63], [64], redes neuronales convolucionales (*Convolutional Neural Networks*, CNNs) [65]-[69] y redes neuronales recurrentes (*Recurrent Neural Networks*, RNNs) [70].

Este tipo de redes aprendían características relevantes en la representación de la información de entrada, pero eran incapaces de capturar dependencias a largo plazo entre las épocas del sueño debido al reducido contexto proporcionado. Esta capacidad resulta fundamental en el rendimiento de la clasificación del sueño, debido a la naturaleza intrínseca en la transición lenta de los procesos fisiológicos que subyacen a las etapas del sueño. En consecuencia, surgieron modelos que entrenaban, por un lado, la codificación independiente de cada época en un vector de características y, en siguiente lugar, una red adicional que recibía como entrada una amplia secuencia que contenía los vectores generados en el paso anterior, lo que se conoce como entrenamiento en dos fases (*two-stage training*).

Las primeras redes híbridas descritas, propuestas por Supratak et al. [71], incrementaron el rendimiento de forma significativa, destacando entre el resto de modelos existentes. No obstante, los primeros trabajos presentados tenían algunas limitaciones, ya que el entrenamiento independiente de las dos subredes en dos etapas no tenía en cuenta el aprendizaje conjunto de la representación de las épocas y el consiguiente modelado secuencial. Además, los modelos eran capaces de clasificar una única época a partir de una secuencia, normalmente la última de la ventana de entrada, comprendiendo un contexto conformado por el lado izquierdo a la época objetivo, aspecto que Olesen et al. en [72] determinaron que resultaba en una reducción de la precisión respecto al uso de un modelo más equilibrado.

Por consiguiente, a partir de estos resultados, desde finales de 2018 se ha experimentado una afluencia de arquitecturas de redes neuronales con capacidad de modelar y capturar dependencias a largo plazo, enmarcadas en un esquema secuencia a secuencia (*sequence-to-sequence*) [73]. En la Figura 2.1 se ilustra esta idea, de forma que la entrada de la red está constituida por  $L$  épocas ( $S_1, \dots, S_L$ ), donde cada elemento  $S_i$ , en el rango  $1 \leq i \leq L$ , puede componerse de datos en crudo o imágenes tiempo-frecuencia, así como pertenecer al registro de uno o más canales de información. En primer lugar, se aplica un codificador de época  $F_E : S \mapsto x$ , obteniendo de cada época  $S$  una representación en forma de vector  $x$ . El resultado de este proceso es una secuencia de vectores  $(x_1, \dots, x_L)$ , representados como círculos verdes. En siguiente lugar, se aplica el codificador de secuencia  $F_S : (x_1, \dots, x_L) \mapsto (z_1, \dots, z_L)$  a la secuencia  $(x_1, \dots, x_L)$  obtenida, generando una nueva secuencia de vectores  $(z_1, \dots, z_L)$ , representados con círculos rojos, que tienen en cuenta tanto las características intrínsecas en cada época, así como el contexto y la información que se obtiene de la ventana de épocas de tamaño  $L$ . Finalmente, la secuencia  $(z_1, \dots, z_L)$  es la que se utiliza para la clasificación y obtención de la etapa del sueño correspondiente.

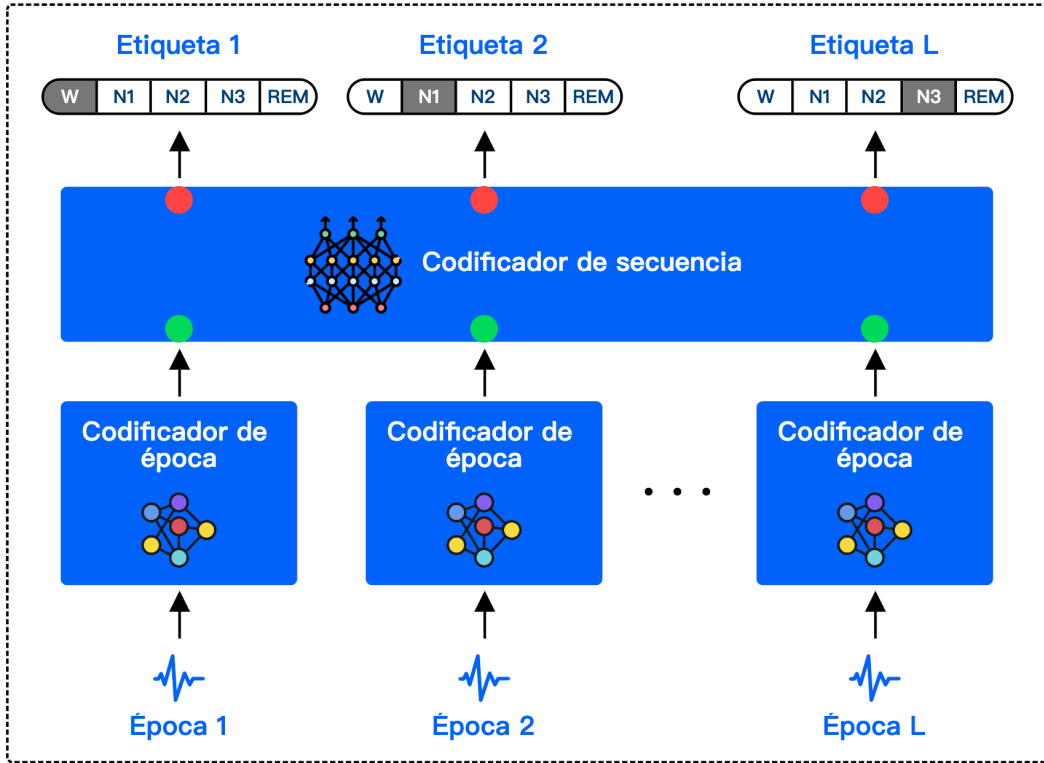


Figura 2.1: Esquema secuencia a secuencia (*sequence-to-sequence*) utilizado en redes neuronales para la clasificación automática del sueño [24].

El esquema secuencia a secuencia supera las limitaciones de los modelos híbridos debido a dos factores principales. Por un lado, las funciones  $F_E$  y  $F_S$  se optimizan al mismo tiempo, permitiendo la interacción entre ambos componentes de la red. Por otro lado, el modelo clasifica todas las épocas de una secuencia de entrada a la vez, de manera que, para cada época, se tiene en cuenta el contexto en función de su posición absoluta dentro de la secuencia, es decir, para la codificación  $x_i$  de una época  $S_i$  el codificador tiene información anterior  $(x_1, \dots, x_{i-1})$  y posterior  $(x_{i+1}, \dots, x_L)$ . El resultado de la aplicación de esta nueva metodología es un incremento en el rendimiento de la evaluación automática de las fases del sueño [74], [75].

La tendencia en los últimos años constituye la aplicación de este patrón sobre diferentes tipos de arquitecturas de modelos de aprendizaje profundo. Entre ellas, destaca la aparición de las redes Transformer, propuestas por Vaswani et al. en [76], basadas en mecanismos de atención, las cuales han tenido un gran éxito en tareas de procesamiento de lenguaje natural (*natural language processing*, NLP) con el desarrollo de modelos como BERT [77] y GPT [78].

La Tabla 2.1 contiene una recopilación de los modelos de aprendizaje profundo más relevantes presentados en los últimos años, en la que se puede observar el año de publicación, la arquitectura de la red y el rendimiento respecto a la métrica de exactitud (*accuracy*) en los datasets SleepEDF [79], [80], MASS [81] y SHHS [82], [83]. Estos valores no justifican la efectividad de un algoritmo sobre otro, sino que pretenden mostrar que obtienen un rendimiento óptimo para los datos, la implementación y la configuración en la que han sido evaluados.

| Modelo                       | Año  | Arquitectura                          | SleepEDF-20 | SleepEDF-78 | MASS | SHHS |
|------------------------------|------|---------------------------------------|-------------|-------------|------|------|
| DeepSleepNet [71]            | 2017 | CNN + RNN (BiLSTM)                    | 82.0        | -           | 86.2 | -    |
| SeqSleepNet [74]             | 2019 | RNN (BiLSTM) + RNN (BiLSTM)           | -           | 82.6        | 82.8 | 86.5 |
| SleepEEGNet [84]             | 2019 | CNN + RNN (BiRNN)                     | 81.5        | -           | -    | -    |
| TinySleepNet [85]            | 2020 | CNN + RNN (LSTM)                      | 85.4        | 83.1        | 87.5 | -    |
| XSleepNet1 [75]              | 2020 | CNN + RNN                             | 82.2        | 80.3        | 85.1 | 87.5 |
| XSleepNet2 [75]              | 2020 | RNN + RNN                             | 83.9        | 80.3        | 85.2 | 87.6 |
| IITNet [86]                  | 2020 | CNN + RNN (BiLSTM)                    | 83.6        | -           | 86.2 | 86.3 |
| AttnSleep [87]               | 2021 | CNN + Multi-Head Self-Attention       | 84.4        | 81.3        | -    | 84.2 |
| AT-BiLSTM [88]               | 2021 | RNN (BiLSTM) + Attention              | 83.78       | -           | -    | -    |
| SleepTransformer [89]        | 2022 | Transformer                           | -           | 81.4        | -    | 87.7 |
| SleepXAI [90]                | 2022 | CNN + CRF                             | 86.2        | -           | -    | -    |
| ENGELBERT [91]               | 2022 | CNN + Transformer                     | 87.1        | -           | 89.2 | -    |
| Cross-Modal Transformer [92] | 2022 | CNN + Transformer                     | -           | 84.0        | -    | -    |
| SAN [93]                     | 2023 | CNN + Multi-Head Self-Attention       | -           | 85.4        | 86.4 | -    |
| TSA-Net [50]                 | 2023 | CNN + Multi-Head Self-Attention + CRF | 86.64       | 82.21       | -    | -    |
| L-SeqSleepNet [94]           | 2023 | RNN (BiLSTM) + Attention              | 86.1        | -           | -    | 88.4 |
| SleepTNT [95]                | 2023 | Transformer                           | -           | 83.1        | -    | -    |

Tabla 2.1: Recopilación de modelos de *deep learning* para la clasificación automática de las etapas del sueño.

Finalmente, cabe mencionar algunas limitaciones que presentan los modelos existentes en el estado del arte. En primer lugar, aunque es cierto que los algoritmos obtienen un rendimiento equiparable al nivel de un técnico especializado en sujetos mayormente sanos, todavía queda avanzar sobre los resultados que arrojan en pacientes con patologías del sueño. Asimismo, muchos algoritmos utilizan varios canales de información (EEG, EOG y EMG), aumentando la complejidad de los dispositivos de medición de señales y la consecuente per-

turbación del sueño de las personas, de manera que el objetivo radica en el uso de un número reducido de electrodos EEG para realizar la clasificación del sueño en entornos domésticos. Además, los algoritmos se entrenan sobre grandes conjuntos de datos que permiten evaluarlos de forma *offline* a través de amplias secuencias. Sin embargo, es necesario analizar su efectividad reduciendo la resolución temporal y sin proporcionar información futura, con el propósito de desarrollar sistemas que funcionen en tiempo real y permitan ampliar el espectro de aplicaciones prácticas de esta tecnología.

En consecuencia, los aspectos descritos proporcionan una justificación para seguir investigando acerca de la clasificación automática del sueño, trabajando en la línea de modelos que entrenen con datos de pacientes más heterogéneos y con un reducido número de canales de información a través de una forma de evaluación *online*.

## 2.2. Explicabilidad e interpretabilidad

La interpretabilidad resulta fundamental en el entendimiento de un sistema de clasificación del sueño debido a la ambigüedad presente en las etapas del sueño, la cual provoca que incluso técnicos especializados difieran en la evaluación manual de una misma grabación [96]. Por consiguiente, la aplicación de métodos de explicabilidad permite revelar el valor clínico existente en las técnicas de clasificación automática del sueño, así como incrementar la fiabilidad de estos sistemas, de forma que se impulse su aceptación y adopción en el sector médico.

Existen algunas aproximaciones de introducción de explicabilidad en sistemas de evaluación del sueño basados en aprendizaje profundo, fundamentados principalmente en métodos de visualización de características, como mapas de sensibilidad [66], mapas de activación de clase ponderados por gradiente guiado (*Guided Grad-CAM*) [97] y mapas de saliencia [98].

Concretamente, Lee et al. [99] propusieron asociar el proceso de aprendizaje a los patrones de las ondas EEG definidos por expertos (descritos en la sección 1.4), utilizando este conocimiento como plantilla para los primeros *kernels* convolucionales de una red CNN, permitiendo localizar estos patrones en la señal EEG mediante la maximización de la similitud del coseno. Asimismo, Dutt et al. [90] aplicaron mapas de activación de clase ponderados por gradiente (*Grad-CAM*) en una red convolucional para ilustrar un mapa de calor sobre la señal original del EEG y encontrar los mismos patrones de activación respecto a las indicaciones propuestas en el manual de la AASM [33].

Por otro lado, Phan et al. [89] realizaron un estudio de la confianza del modelo utilizando las probabilidades de salida para determinar si los fallos se concentraban en épocas clasificadas con alta incertidumbre. También, utilizaron los pesos de autoatención para mostrarlos en un mapa de calor sobre la señal original, además de conseguir aplicarlos con el propósito de ilustrar la influencia de las épocas vecinas en la clasificación de la época objetivo, y fueron capaces de reconstruir la señal de entrada a partir de las características aprendidas para compararla con la original. Estas técnicas se fundamentan en los mecanismos de atención correspondientes a la arquitectura de una red Transformer, mientras que las estrategias comentadas en el párrafo anterior se apoyan en las propiedades de una red convolucional.

Sin embargo, la mayoría de modelos existentes no aplica metodologías de interpretabilidad que posibiliten un mayor entendimiento acerca de su funcionamiento. Además, sigue siendo una incógnita cómo cuantificar, evaluar y comparar objetivamente la explicabilidad de un

sistema. A pesar de todo el progreso en el campo, la evaluación automática de las fases del sueño no ha sido ampliamente adoptada en el ámbito clínico. Por lo tanto, el presente proyecto se encuentra ante un contexto en el que la investigación sobre este tipo de técnicas adquiere una relevancia crucial, estableciendo como objetivo el desarrollo y validación de métodos de interpretación que contribuyan a la adopción de modelos de aprendizaje profundo para la clasificación automática de las etapas del sueño, los cuales permitan que los profesionales médicos dejen de considerar las herramientas de inteligencia artificial como una caja negra y posibiliten el cumplimiento de los mecanismos de regulación y legislación requeridos.

## Capítulo 3

# Metodología

Una vez completada la revisión de los métodos existentes para la evaluación automática del sueño, es evidente que los resultados obtenidos mediante el uso de técnicas de aprendizaje profundo son los más prometedores hasta el momento. Por lo tanto, este capítulo describe la arquitectura del modelo de *deep learning* implementado para la clasificación automática del sueño, las técnicas de explicabilidad e interpretabilidad aplicadas sobre este y la metodología de evaluación del sistema desarrollado.

### 3.1. Arquitectura del modelo

La arquitectura desarrollada combina redes convolucionales con las emergentes redes Transformer, propuestas por Vaswani et al. en [76], y está basada en la red presentada por Brandmay et al. en [91]. El esquema del sistema de clasificación automática del sueño se ilustra en la Figura 3.1, y se compone de un primer bloque convolucional (CNN) encargado de la codificación a nivel de época (extracción de características) y un segundo componente Transformer que realiza la codificación de la secuencia obtenida. En cuanto a aspectos generales, el modelo utiliza como entrada la señal en bruto, a diferencia de otras redes que trabajan con imágenes tiempo-frecuencia. Además, la optimización del conjunto de parámetros ( $|\theta| = 737.605$ ) se ha llevado a cabo mediante descenso de gradiente estocástico (*Stochastic Gradient Descent*, SGD) con el algoritmo AdamW [100] ( $learning\ rate = 10^{-3}$ ,  $weight\ decay = 10^{-4}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  y  $\epsilon = 10^{-8}$ ), el cual minimiza la función de pérdida de entropía cruzada categórica (*categorical cross-entropy*) debido a la existencia de más de 2 clases en la clasificación.

La primera parte de la red se corresponde con los codificadores de época, representados en la Figura 3.1a en la entrada del esquema, que se encargan de codificar cada elemento  $S_i$ , en el rango  $1 \leq i \leq L$ , de la secuencia de entrada  $(S_1, \dots, S_L)$ , la cual comprende la dimensión  $(L, \text{núm.muestras}, \text{núm.canales})$ , en una secuencia de vectores de características  $(x_1, \dots, x_L)$ , donde  $L$  es el tamaño de la ventana de épocas del sueño. Esta secuencia de vectores se obtiene aplicando a cada época de la secuencia de entrada una red neuronal convolucional (CNN) para extraer las características representativas de la señal en cuestión.

Los aspectos más relevantes de la subred están representados en la Figura 3.1b, la cual se compone principalmente de sucesivas capas convolucionales 1D, cada una con 64 filtros, un *kernel* de tamaño igual a 5, un valor de *stride* de 1 y la función de activación ReLU. También, se han incluido capas de normalización por lotes (Batch Normalization) [101] después de cada convolución, así como capas de Spatial Dropout 1D y Dropout [102] a modo de regularización, con ratios de 0.1 y 0.05, respectivamente. En algunos bloques convolucionales intermedios se han aplicado operaciones de Max Pooling 1D, con un tamaño y *stride* de 2,

mientras que la última operación es un Global Max Pooling 1D, la cual reduce la dimensión de los datos. Cabe destacar que el parámetro  $N_B = 4$  se corresponde con el número total de los bloques indicados que se concatenan. Como resultado del procedimiento descrito, se obtiene, para cada época, un vector de características de tamaño 64, de forma que la secuencia de vectores  $(x_1, \dots, x_L)$  resultante para una ventana de épocas de entrada  $(S_1, \dots, S_L)$  de tamaño  $L$  comprende una matriz de dimensión  $(L, 64)$ .

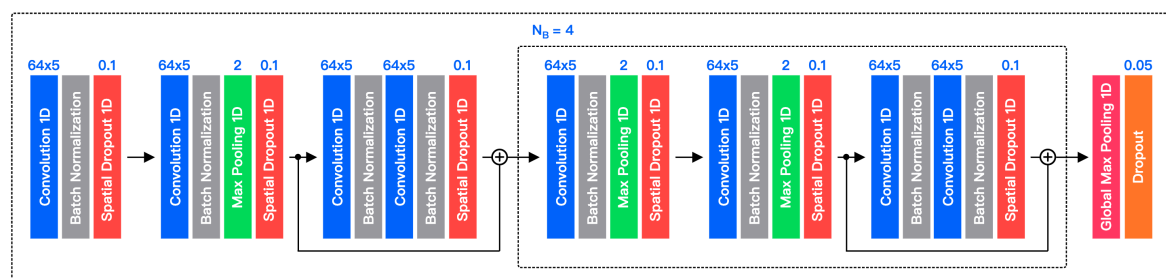
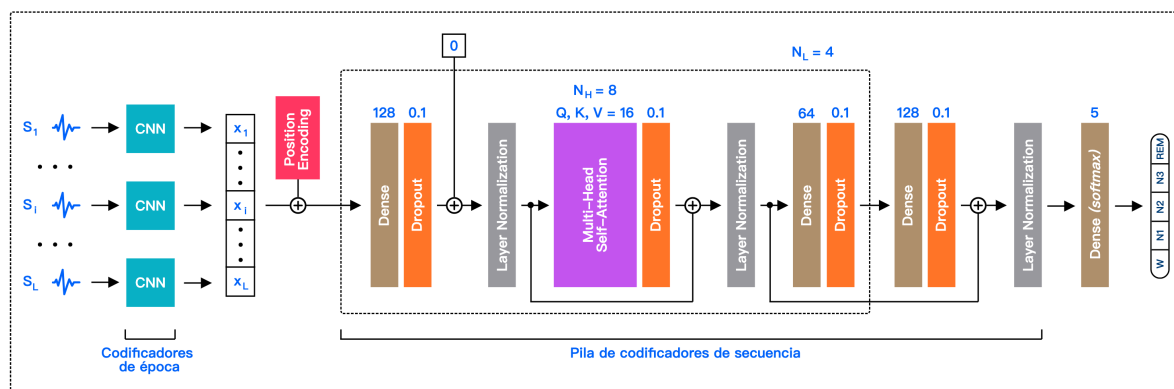


Figura 3.1: Arquitectura del modelo de aprendizaje profundo.

del sueño, se añade información sobre la posición de los tokens en la secuencia. Para ello, tal y como se propone en [76], el módulo *Position Encoding*, presentado en la Figura 3.1a, suma a la secuencia de vectores  $(x_1, \dots, x_L)$  de dimensión  $(L, 64)$  una matriz determinista de codificación posicional (*positional encodings*) [103] del mismo tamaño, la cual se obtiene siguiendo las ecuaciones 3.1 y 3.2:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{(2i/d_{model})}}\right) \quad (3.1)$$

$$PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{(2i/d_{model})}}\right) \quad (3.2)$$

donde  $pos$  (fila) es el índice de la posición de cada elemento de la secuencia de entrada en el rango  $0 \leq pos < L$ , mientras que  $i$  (columna) comprende el espectro de valores asociados a la mitad de la dimensión del modelo ( $d_{model} = 64$ ), en este caso  $0 \leq i < \frac{64}{2}$ , de manera que en las columnas pares se aplica la función del seno y en las impares la del coseno.

Es decir, cada dimensión de la codificación posicional se corresponde con una senoide, donde las longitudes de onda forman una progresión geométrica de  $2\pi$  a  $10000 \cdot 2\pi$ . Esta función permite que el modelo aprenda la atención por posiciones relativas, ya que para cualquier desplazamiento fijo  $k$ ,  $PE_{pos+k}$  puede representarse como una función lineal de  $PE_{pos}$ . Alternativamente, existe otra forma de codificación posicional en la que la matriz no es determinista, sino que se aprende como un elemento más del entrenamiento del modelo, la cual también ha sido implementada y probada. Sin embargo, a través del experimento descrito en el Apéndice B, se ha determinado que los resultados arrojados por la codificación determinista son ligeramente superiores a esta opción, resultando en la técnica desarrollada.

Posteriormente, el resultado de este proceso constituye la entrada de los codificadores Transformer, donde el parámetro  $N_L = 4$  define el número total de bloques apilados. Cada uno de estos componentes consta de una capa densa con 128 neuronas, que realiza una expansión respecto a la dimensión de entrada. A la salida producida se le suma la salida del bloque anterior, de forma que al primer bloque se le añade el valor 0, tal y como se indica en la Figura 3.1a, todo ello seguido de una capa de normalización por lotes. A continuación, se aplica una capa Multi-Head Self-Attention, imprescindible en este tipo de arquitectura, en la que se ha definido el número de cabezas de atención  $N_H = 8$  y la consecuente dimensión de las variables *query*, *key* y *value* como  $d_{model}/N_H = 128/8 = 16$ , donde cabe mencionar que  $d_{model} = 128$  debido a la expansión realizada. El *encoder* termina con otra capa de normalización por lotes y una compresión de la dimensión mediante una capa densa compuesta de 64 neuronas.

Después del último bloque de la pila se lleva a cabo una última expansión con una capa densa de 128 neuronas, cuya salida atraviesa una capa de normalización por lotes. Finalmente, la capa densa que cuenta con 5 neuronas y una función de activación *softmax* produce las probabilidades de salida asociadas a cada etapa del sueño. Cabe destacar que se han aplicado capas de Dropout con un ratio de 0.1 después del módulo *Position Encoding* y a continuación de cada capa densa y Multi-Head Self-Attention, junto con varias operaciones de suma sobre la salida de una capa respecto a la salida de capas anteriores (ver Figura 3.1a).

En definitiva, el modelo desarrollado se compone de una primera parte convolucional encargada de extraer las características intrínsecas de la señal de entrada, de forma independiente para cada elemento de la secuencia, y un segundo componente Transformer que trabaja a nivel de codificación de secuencia. En el Apéndice C se describe un estudio de ablación (*ablation*) en el que se experimenta con el valor de los hiperparámetros de la red. Adicionalmente, se han realizado algunas pruebas con el modelo SleepXAI [90], las cuales se detallan en el Apéndice D. Sin embargo, el presente trabajo se centra en la arquitectura descrita en esta sección, tratándose de la propuesta definitiva en cuanto al modelo de aprendizaje profundo utilizado para la clasificación automática del sueño. Para más detalles sobre las herramientas utilizadas en la implementación del sistema, ver Apéndice A.

## 3.2. Técnicas de explicabilidad e interpretabilidad

El objetivo de esta sección consiste en describir el conjunto de técnicas de explicabilidad e interpretabilidad que se han aplicado sobre el modelo desarrollado, descrito en la sección anterior, de manera que se posibilite la validación de múltiples métodos de interpretación y se determine su contribución en cuanto al entendimiento de la inteligencia artificial para su adopción en el ámbito clínico.

### 3.2.1. Cuantificación de la confianza

Al tratarse de un problema de clasificación multiclase, la red neuronal profunda genera un vector  $\hat{y}$  cuyos elementos son probabilidades, las cuales se asocian con cada una de las  $C$  clases (en este caso  $C = 5$ ). En consecuencia, la etapa del sueño a la que le corresponde un valor más alto se considera la predicción del modelo. Sin embargo, esta métrica no indica el nivel de confianza de la red respecto a la decisión tomada. Por ello, cuando el sistema asigna una probabilidad igual a 1 a una clase y 0 al resto, es previsible que el modelo haya llevado a cabo la clasificación de la época con un nivel de confianza alto. Sin embargo, si todos los valores generados para cada clase son idénticos, el sistema presenta incertidumbre en su decisión.

Por consiguiente, Phan et al. [89] utilizaron la entropía de la distribución de probabilidad discreta, entendida como una medida teórica que cuantifica numéricamente la incertidumbre a través de la ecuación que calcula la entropía normalizada:

$$H(\hat{y}) = - \sum_{c=1}^C \hat{y}_c \frac{\log(\hat{y}_c)}{\log C} \quad (3.3)$$

donde el rango de valores se corresponde con  $[0.0, 1.0]$ , asumiendo que  $0 \cdot \log 0 = 0$ . Finalmente, la confianza del modelo se cuantifica del siguiente modo:

$$Conf(\hat{y}) = 1 - H(\hat{y}) \quad (3.4)$$

Como resultado, se obtiene que  $H(\hat{y}) = 1$  y  $Conf(\hat{y}) = 0$  cuando  $\hat{y} = (\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ , mientras que  $H(\hat{y}) = 0$  y  $Conf(\hat{y}) = 1$  cuando  $\hat{y} = (1, 0, 0, 0, 0)$ . De esta forma, se puede determinar si los fallos se concentran en épocas clasificadas con alta incertidumbre, realizando un filtrado a través de un umbral establecido por el que cualquier valor de  $Conf(\hat{y})$  por debajo

de este se asocie con una época de baja confianza. Además, es posible ilustrar la evolución de la confianza de la red y los errores cometidos a lo largo de las diferentes épocas de la noche en un diagrama, junto con los hipnogramas correspondientes al proceso de clasificación del sueño, con el propósito de llevar a cabo un análisis visual de la incertidumbre del sistema.

### 3.2.2. Análisis probabilístico

A partir de las probabilidades de salida del modelo, asociadas a cada una de las 5 etapas del sueño, se va a realizar un estudio de la distribución de estos valores a lo largo de la noche, de forma que se determine si los resultados arrojados se asocian con el hipnograma de la grabación y si existe correspondencia con la transición lenta presente en las fases del sueño. Cabe mencionar que, en contraste con el método anterior, el cual estudia la confianza del modelo a partir de la salida generada, se ha considerado de relevancia diferenciar una técnica que se centre únicamente en analizar los valores de las probabilidades obtenidas, ya que aunque estos valores guardan una relación directa con la confianza del sistema, se pueden extraer conclusiones adicionales y complementarias sobre el funcionamiento del modelo presentado.

Para ello, tal y como se propone en [89], es posible generar un gráfico de barras con la intensidad de las probabilidades en el rango  $[0.0, 1.0]$ , asociadas a las fases del sueño, para cada época contenida en la secuencia total. Por otro lado, también se puede generar un mapa de calor que representa, para cada época, la concentración de las probabilidades asociadas a cada fase del sueño, obteniendo un espectro completo de la distribución de la intensidad de estos valores a lo largo de la grabación correspondiente.

### 3.2.3. Influencia de las épocas vecinas

Cuando un técnico se enfrenta a la clasificación de una época que no muestra suficiente evidencia en cuanto a las características propias de cada fase del sueño, atiende a las épocas vecinas de la época objetivo, las cuales proporcionan un contexto y apoyo considerable para su evaluación [33]. En consecuencia, resulta de gran importancia conocer a qué épocas vecinas de la secuencia de entrada presta mayor atención el modelo cuando clasifica la época objetivo.

Con este propósito, se van a utilizar los pesos de autoatención generados en las capas Multi-Head Self-Attention, pertenecientes al componente Transformer del modelo presentado en la sección 3.1, encargado de la codificación secuencial, para representar en un diagrama de barras cómo afecta el contexto de una época en su evaluación según estos valores. Cada una de estas capas genera una matriz  $A$ , en la que el elemento  $a_{ij}$  indica la medida en la que el peso de autoatención  $j$  contribuye a la representación de la época objetivo  $i$ . La dimensión de esta matriz comprende el rango  $(L, 128)$ , donde  $L$  es el tamaño de la secuencia de entrada y el valor 128 representa la dimensión de la codificación de las épocas en forma de pesos de autoatención.

A partir de este planteamiento, mediante el producto  $A[i] \cdot A^T$  para la época objetivo  $i$ , se obtiene un vector de tamaño  $L$ , cuyos valores pueden ser normalizados y mostrados sobre el rango  $[1, L]$  en un gráfico de barras, de manera que se ilustre la contribución de los pesos de autoatención de las épocas vecinas en la evaluación de la época objetivo  $i$ . Este proceso puede repetirse para cada época de la secuencia de entrada  $(S_1, \dots, S_L)$ , en el rango  $1 \leq i \leq L$ , obteniendo el espectro completo. La metodología descrita está inspirada en los trabajos de Phan et al. [89] y Pradeepkumar et al. [92].

### 3.2.4. Grad-CAM

El método Grad-CAM [104] es una técnica de interpretación visual que permite comprender cómo una red neuronal convolucional toma decisiones. En el campo de la visión por computador, el objetivo consiste en identificar las regiones de una imagen que resultan importantes para la clasificación, lo que se traduce en localizar las características visuales a las que atiende el modelo. Técnicamente, esta metodología utiliza los gradientes de la función de pérdida respecto a las activaciones de las capas convolucionales de la red, los cuales se ponderan y promedian espacialmente para obtener un mapa de activación que resalta las zonas clave de la imagen original.

El proceso de clasificación de las etapas del sueño que realizan los profesionales sanitarios requiere del análisis visual de los datos recopilados. Por ello, surge la posibilidad de aplicar el método Grad-CAM sobre series temporales en lugar de imágenes, identificando las zonas de la señal que son más relevantes para la decisión del modelo. De esta forma, es posible ilustrar un mapa de calor sobre la señal original del EEG para encontrar los mismos patrones de activación respecto a las indicaciones que se describen en el manual de la AASM [33].

Este es el procedimiento que proponen Dutt et al. [90] para el modelo que presentan en su artículo, y es el que se ha aplicado sobre la red convolucional (Figura 3.1b) del sistema de aprendizaje profundo desarrollado en el presente trabajo. Por consiguiente, se aporta explicabilidad en la codificación a nivel de época, a diferencia de la técnica anterior, la cual se centra en la interpretabilidad de la codificación a nivel secuencial, correspondiente al módulo Transformer. A grandes rasgos, los pasos que se han seguido para la implementación del método son los siguientes:

1. Definir la capa convolucional de la que se va a obtener la salida.
2. Obtener el modelo hasta la capa convolucional escogida, manteniendo las capas anteriores intactas.
3. A partir de una determinada época de entrada, calcular la salida con el modelo convolucional, generado en el paso anterior, y con el modelo completo.
4. Obtener los gradientes respecto a la salida del modelo convolucional en base al valor de la probabilidad asociada a la predicción de la clase realizada por el modelo completo, los cuales tienen una dimensión  $(n, \text{núm\_filtros})$ , donde  $n$  depende de la capa convolucional seleccionada.
5. Computar la media de los gradientes reduciendo el eje correspondiente a la dimensión  $n$ , obteniendo una media de la intensidad de los gradientes para cada característica extraída por la capa convolucional elegida, generando un vector de tamaño  $\text{núm\_filtros}$ .
6. Sumar el vector de intensidad media a la salida del modelo convolucional, de manera que se añada información sobre cómo de importante es cada canal de la capa convolucional para la predicción realizada por el modelo completo.
7. Calcular la media reduciendo el eje correspondiente a la dimensión  $\text{núm\_filtros}$ , obteniendo un vector de tamaño  $\text{núm\_muestras}$  que puede ser mostrado en forma de mapa de calor sobre la señal original, el cual contiene los valores de activación de la capa convolucional respecto a la predicción del modelo para cada instante temporal.

### 3.3. Evaluación

La presente sección detalla el conjunto de datos disponibles para la ejecución del modelo de aprendizaje profundo presentado en la sección 3.1, junto con la metodología de evaluación del rendimiento arrojado sobre estos datasets del método propuesto y desarrollado, métricas computadas y pruebas realizadas.

#### 3.3.1. Datos disponibles

Los datos disponibles para la ejecución del sistema se dividen en dos clases: datasets públicos, que permiten comparar el rendimiento con los modelos existentes en el estado del arte, y datasets recopilados en los laboratorios de Bitbrain, los cuales posibilitan evaluar la aplicación directa del algoritmo en un contexto práctico, realista y controlado.

##### Datasets públicos

Los datasets públicos utilizados para la evaluación del modelo de *deep learning* en la tarea de la clasificación del sueño son los siguientes:

- **SleepEDF** [79], [80]: contiene un conjunto de sujetos asociado al estudio del efecto de la edad en personas sanas (SC), y un segundo grupo perteneciente al estudio de los efectos del temazepam en el sueño (ST). En este proyecto se utilizan los 20 sujetos del conjunto SC (edad  $28.9 \pm 2.9$ ), en el que cada registro del PSG se compone de 2 canales EEG (Fpz-Cz y Pz-Oz) con una frecuencia de muestreo de 100 Hz, sin la aplicación de ningún procesamiento o filtrado. La clasificación de los datos fue realizada manualmente por expertos siguiendo el estándar de R&K [32], utilizando ocho clases (W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN). Sin embargo, las etapas N3 y N4 se fusionaron en una única fase N3, mientras que las etapas MOVEMENT y UNKNOWN fueron eliminadas, todo ello para seguir el estándar de la AASM [33]. Además, los periodos largos de vigilia (W) en los que el sujeto no duerme fueron excluidos, incluyendo únicamente los 30 minutos previos y posteriores al periodo del sueño.
- **Dreem Open Dataset** [105]: repositorio de datos sobre el sueño constituido por 2 conjuntos. El primero se denomina *Dreem Open Dataset - Healthy* (DOD-H), compuesto por 25 sujetos sanos, y el segundo se identifica como *Dreem Open Dataset - Obstructive* (DOD-O), formado por 50 sujetos sospechosos de padecer trastornos respiratorios relacionados con el sueño. Contiene información sobre el EOG, EMG, EKG y EEG (C3-M2, F4-M1, F3-F4, F3-M2, F4-O2, F3-O1, Fp1-F3, Fp1-M2, Fp1-O1, Fp2-F4, Fp2-M1, Fp2-O2) muestreados a 250 Hz y clasificados por 5 técnicos siguiendo el manual de la AASM.
- **ISRUC-Sleep** [106]: conjunto de datos de dominio público compuesto por 3 subgrupos (I-III). Los subconjuntos I y II están formados por 100 y 8 grabaciones del PSG, respectivamente, de adultos con indicios de trastornos del sueño, mientras que el subgrupo III consta de 10 sujetos sanos. Los datos recopilados contienen información sobre el EOG, EEG (F3-A2, F4-A1, C3-A2, C4-A1, O1-A2, O2-A1), EMG, EKG, SaO2, posición, flujo de aire y esfuerzo abdominal, todo ello muestreado a 200 Hz. La evaluación de las fases del sueño fue completada por 2 expertos según las directrices de la AASM. Debido al desequilibrio en el número de sujetos entre subconjuntos, se van a utilizar únicamente los datos del subgrupo I.

Cabe destacar que los datos pertenecientes a Dreem Open Dataset e ISRUC-Sleep han sido remuestreados a 100 Hz y procesados con un filtro paso banda entre 0.5 y 30 Hz, junto con un filtro *notch* a 50 Hz, con el objetivo de compensar la baja relación señal-ruido (*signal-to-noise ratio*, SNR) existente en los sistemas EEG. En cambio, los sujetos del dataset SleepEDF se han mantenido intactos, siguiendo la misma convención presente en la literatura respecto a este dataset.

## Datasets de Bitbrain

Por otro lado, los datos recopilados por Bitbrain se obtienen en un laboratorio del sueño totalmente equipado, el cual cuenta con un polisomnógrafo completo de calidad médica que actúa como dispositivo de referencia. Además, la grabación se realiza de forma simultánea con la diadema textil [25] con la que se pretende integrar el modelo de aprendizaje profundo desarrollado. En consecuencia, se dispone de la siguiente información:

- 6 canales EEG del dispositivo PSG (F3, F4, C3, C4, O1, O2).
- 2 canales EOG que miden el movimiento vertical ocular (EOGV) y el movimiento horizontal ocular (EOGH).
- 1 canal EMG que registra la actividad muscular de la cara.
- Información sobre la saturación de oxígeno en sangre medida con un pulsioxímetro colocado en el dedo.
- 5 canales EEG de la banda textil (AF7, AF8, Fp1, Fp2, T8).

Las grabaciones fueron muestreadas a 256 Hz y remuestreadas a 128 Hz. Además, se aplicó un filtro paso banda entre 0.5 y 45 Hz. A partir de este planteamiento, se cuenta con dos estudios (*Study 1* y *Study 3*) con 30 y 57 sujetos, respectivamente, que recopilan la información descrita. La clasificación del sueño ha sido completada por un único técnico siguiendo las reglas de la AASM. Sin embargo, existe una variante del *Study 1* en el que la evaluación es realizada por 3 expertos, la cual también será utilizada para estudiar el efecto del número de profesionales que puntúan los datos. Finalmente, a modo de resumen, en la Figura 3.2 se ilustran los datasets descritos en esta sección, seleccionando aquellos subgrupos que van a ser utilizados en la evaluación del sueño.

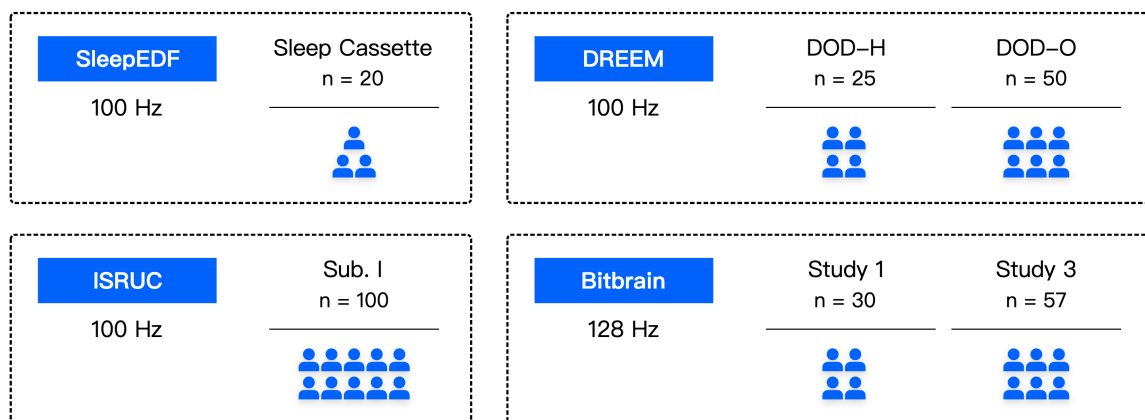


Figura 3.2: Datasets disponibles para la evaluación del modelo de clasificación del sueño.

### 3.3.2. Rendimiento

La metodología implementada para la evaluación del sistema de clasificación del sueño sigue el esquema ilustrado en la Figura 3.3. En primer lugar, se aplica una determinada técnica de filtrado sobre los datos del PSG/EEG en bruto, junto con un remuestreo de la frecuencia de la señal, siempre y cuando el dataset precise de su uso. Además, se lleva a cabo un escalado en forma de matriz 3D, cuyas dimensiones se corresponden con el número de épocas, el número de muestras o puntos temporales (definido como el producto  $frecuencia \cdot 30$ ) y el número de canales empleados.

En siguiente lugar, los datos se dividen en entrenamiento y test a través de una proporción del 90% y el 10%, respectivamente, con el objetivo de llevar a cabo un *10-Fold Cross-Validation*. A su vez, los datos de entrenamiento se dividen en un subconjunto de entrenamiento (85%) y un subconjunto de validación (15%). A continuación, se aplica la normalización Z-Score a las grabaciones, utilizando únicamente los datos de entrenamiento. El resultado de este proceso conforma la entrada del modelo de aprendizaje profundo.

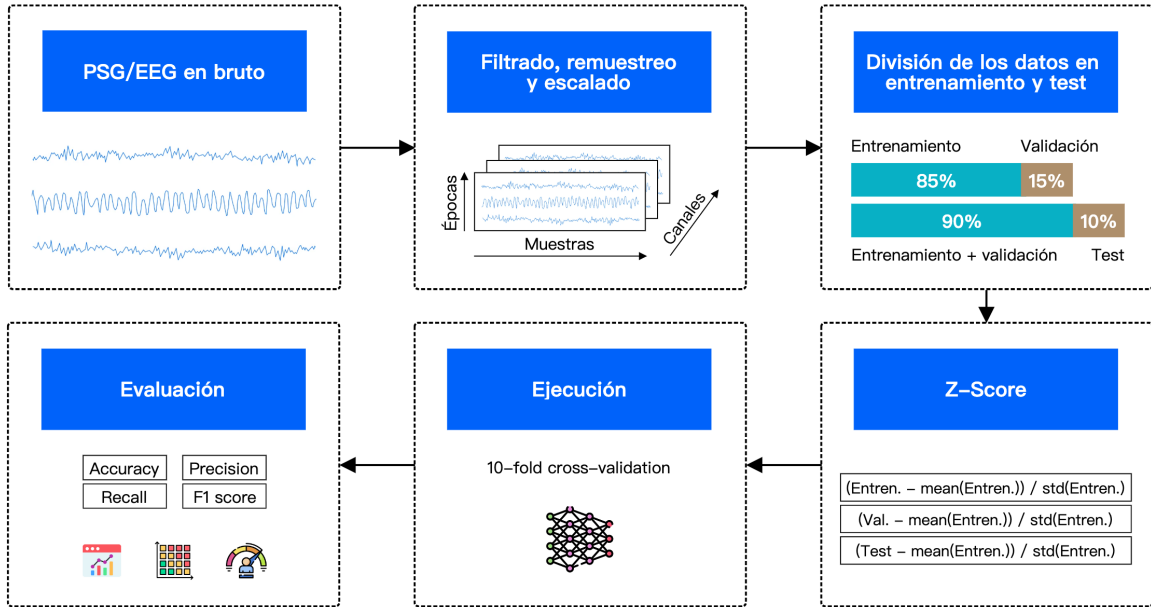


Figura 3.3: Esquema de los pasos para la evaluación del método propuesto.

El rendimiento del modelo se evalúa a través de la matriz de confusión generada y un informe de clasificación que mide el nivel de acuerdo entre las etiquetas de los expertos y del algoritmo. La primera métrica del informe se corresponde con la exactitud o *accuracy* (Ecuación 3.5), la cual calcula el número de predicciones correctas dividido por el número total de predicciones. Concretamente, se define como la suma de verdaderos positivos (TP) y verdaderos negativos (TN) entre la suma de los TP, TN, falsos positivos (FP) y falsos negativos (FN). También, se computa el valor de la precisión (*precision*) como el cociente de los TP y el total de ejemplos clasificados como positivos (Ecuación 3.6). Por último, se calcula la exhaustividad o *recall* (Ecuación 3.7), la cual define la tasa de verdaderos positivos, y la medida F1 (*F1 score*), que se corresponde con la media armónica de la precisión y la exhaustividad (Ecuación 3.8). Por simplicidad, en el resto del presente documento se van a utilizar los términos definidos en su forma original en inglés.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.5)$$

$$Precision = \frac{TP}{TP + FP} \quad (3.6)$$

$$Recall = \frac{TP}{TP + FN} \quad (3.7)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3.8)$$

En cuanto a los parámetros de ejecución para el entrenamiento del modelo, se ha definido el tamaño de la secuencia de entrada  $L = 100$  (*batch\_size*), el número de épocas de entrenamiento en cada *fold* con el valor 150, el número de pasos por época de entrenamiento como el cociente  $|épocas\_entrenamiento| / L$ , y el número de pasos para la validación al finalizar cada época de entrenamiento como  $|épocas\_validación| / L$ .

Debido a que muchas de las pruebas están diseñadas para evaluar el modelo en tiempo real (*online*), cumpliendo con uno de los objetivos planteados al comienzo del trabajo, ha sido necesario implementar un mecanismo que permita llevar a cabo la clasificación del sueño sin la existencia de información temporal futura. Para ello, se ha desarrollado una técnica que simula la evaluación en línea, la cual alimenta la red a través de una ventana temporal deslizante de 2.5 minutos (5 épocas de 30 segundos), de forma que cada época objetivo se concatena al final de esta, y la época más antigua se descarta, siguiendo una política FIFO. En consecuencia, únicamente se cuenta con información temporal acerca del pasado, útil para el contexto que aportan las épocas vecinas de la época objetivo. Adicionalmente, se plantea una forma de evaluación *offline* o causal (con información pasada y futura) como referencia para estudiar el impacto del contexto temporal en el rendimiento de una evaluación puramente *online*, la cual ha sido implementada a través de ventanas temporales no solapadas del mismo tamaño  $L$  que la secuencia de entrada, en las que todas las épocas se etiquetan al mismo tiempo, a diferencia de la técnica de clasificación *online*, en la que se puntúan una por una.

Una vez establecidos los aspectos de configuración para la evaluación del rendimiento del sistema, en la Tabla 3.2 se expone el conjunto de pruebas planteadas sobre los datasets públicos disponibles, mientras que en la Tabla 3.1 se describen los tests a ejecutar con los datos de Bitbrain.

En primer lugar, el dataset público SleepEDF se va a utilizar como referencia para la comparativa del rendimiento del modelo respecto a los resultados expuestos en la literatura científica, ya que se trata de uno de los conjuntos de datos más empleado. Cabe mencionar que, a diferencia del esquema presentado en la Figura 3.3 y aplicado en el resto de tests, la metodología de evaluación ha sido ligeramente modificada, de forma que se lleva a cabo un *Leave-One-Out Cross-Validation* en lugar de un *10-Fold Cross-Validation*, ya que constituye la técnica desarrollada en la mayoría de los artículos que utilizan este dataset. Teniendo esto en cuenta, se han planteado un total de 4 pruebas, en las que se evalúa de forma *offline* y *online* la información de los canales Fpz-Cz y Pz-Oz.

En siguiente lugar, para los datasets DOD-H, DOD-O e ISRUC-Sleep se proponen 3 pruebas sobre un único canal EEG, correspondiente a la región frontal (F3), central (C3) y occipital (O1), respectivamente, con el propósito de comprender el papel que desempeñan las posiciones de los electrodos en la clasificación del sueño. También, se exponen dos tests en los que se explora el impacto de combinar varias fuentes de información, utilizando únicamente el EEG (regiones frontal, central y occipital) y mezclando el EEG, EMG, EKG y EOG. Finalmente, se plantea una prueba *offline* de referencia sobre la región frontal (F3).

Por último, para el primer dataset de Bitbrain (*Study 1*), se han propuesto 4 tests sobre los datos del PSG, en los que se realiza una evaluación *offline* y *online* en los canales F3 + F4, utilizando tanto los datos puntuados por un único experto como el conjunto evaluado por 3 técnicos. Respecto al dataset *Study 3*, se exponen 2 pruebas con los datos del PSG en los canales F3 + F4, y 2 experimentos con los datos recopilados mediante la diadema textil en los canales AF7 + AF8 (EEG), todos ellos a través de una evaluación *offline* y *online*. Esta última prueba resulta de gran relevancia para el estudio, ya que utiliza las grabaciones de la banda textil con la que se pretende integrar el modelo. Cabe destacar que, a diferencia de los tests anteriores, se han realizado algunas modificaciones en los parámetros, debido a que se han encontrado valores óptimos diferentes para estos conjuntos en cuanto al número de épocas de entrenamiento para cada *fold* y el tamaño de la secuencia de entrada  $L$  (ver Apéndice C), especificados en la Tabla 3.1 en el caso de sufrir cambios.

|                | Motivación                                                                                                                              | Descripción                                                          |
|----------------|-----------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| <b>Study 1</b> |                                                                                                                                         |                                                                      |
| <i>STD100</i>  | Test offline con información pasada y futura sobre los datos del PSG (referencia) con un único experto en la puntuación de la grabación | Entrenamiento y test en F3 + F4, número de épocas = 200              |
| <i>STD101</i>  | Test offline con información pasada y futura sobre los datos del PSG con el consenso de 3 expertos en la puntuación de la grabación     | Entrenamiento y test en F3 + F4                                      |
| <i>STD102</i>  | Evaluación online sobre los datos del PSG con un único experto en la puntuación de la grabación                                         | Entrenamiento y test en F3 + F4, número de épocas = 200              |
| <i>STD103</i>  | Evaluación online sobre los datos del PSG con el consenso de 3 expertos en la puntuación de la grabación                                | Entrenamiento y test en F3 + F4                                      |
| <b>Study 3</b> |                                                                                                                                         |                                                                      |
| <i>STD300</i>  | Test offline con información pasada y futura sobre los datos del PSG (referencia)                                                       | Entrenamiento y test en F3 + F4, número de épocas = 200, $L = 175$   |
| <i>STD301</i>  | Test offline con información pasada y futura sobre los datos de la banda textil de Bitbrain                                             | Entrenamiento y test en AF7 + AF8, número de épocas = 200, $L = 200$ |
| <i>STD302</i>  | Evaluación online sobre los datos del PSG                                                                                               | Entrenamiento y test en F3 + F4, número de épocas = 200, $L = 175$   |
| <i>STD303</i>  | Evaluación online sobre los datos de la banda textil de Bitbrain                                                                        | Entrenamiento y test en AF7 + AF8, número de épocas = 200, $L = 200$ |

Tabla 3.1: Resumen de las pruebas planteadas sobre los datasets de Bitbrain.

|                    | Motivación                                                                                            | Descripción                                                         |
|--------------------|-------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| <b>SleepEDF</b>    |                                                                                                       |                                                                     |
| <i>SEDF00</i>      | Test offline con información pasada y futura                                                          | Entrenamiento y test en Fpz-Cz                                      |
| <i>SEDF01</i>      | Test offline con información pasada y futura                                                          | Entrenamiento y test en Pz-Oz                                       |
| <i>SEDF02</i>      | Evaluación online sin información futura                                                              | Entrenamiento y test en Fpz-Cz                                      |
| <i>SEDF03</i>      | Evaluación online sin información futura                                                              | Entrenamiento y test en Pz-Oz                                       |
| <b>DOD-H</b>       |                                                                                                       |                                                                     |
| <i>DODH00</i>      | Test offline con información pasada y futura                                                          | Entrenamiento y test en F3-M2                                       |
| <i>DODH01</i>      | Evaluación online con un electrodo frontal                                                            | Entrenamiento y test en F3-M2                                       |
| <i>DODH02</i>      | Evaluación online con un electrodo central                                                            | Entrenamiento y test en C3-M2                                       |
| <i>DODH03</i>      | Evaluación online con un electrodo occipital                                                          | Entrenamiento y test en O1-M2                                       |
| <i>DODH04</i>      | Evaluación online combinando varios canales EEG de la región frontal, central y occipital             | Entrenamiento y test en F3-M2 + C3-M2 + O1-M2                       |
| <i>DODH05</i>      | Evaluación online sin información futura combinando varios canales EEG y utilizando el EMG, EKG y EOG | Entrenamiento y test en F3-M2 + C3-M2 + O1-M2 + EMG + EKG + EOG (2) |
| <b>DOD-O</b>       |                                                                                                       |                                                                     |
| <i>DODO00</i>      | Test offline con información pasada y futura                                                          | Entrenamiento y test en F3-M2                                       |
| <i>DODO01</i>      | Evaluación online con un electrodo frontal                                                            | Entrenamiento y test en F3-M2                                       |
| <i>DODO02</i>      | Evaluación online con un electrodo central                                                            | Entrenamiento y test en C3-M2                                       |
| <i>DODO03</i>      | Evaluación online con un electrodo occipital                                                          | Entrenamiento y test en O1-M2                                       |
| <i>DODO04</i>      | Evaluación online combinando varios canales EEG de la región frontal, central y occipital             | Entrenamiento y test en F3-M2 + C3-M2 + O1-M2                       |
| <i>DODO05</i>      | Evaluación online sin información futura combinando varios canales EEG y utilizando el EMG, EKG y EOG | Entrenamiento y test en F3-M2 + C3-M2 + O1-M2 + EMG + EKG + EOG (2) |
| <b>ISRUC-Sleep</b> |                                                                                                       |                                                                     |
| <i>ISRUC00</i>     | Test offline con información pasada y futura                                                          | Entrenamiento y test en F3-A2                                       |
| <i>ISRUC01</i>     | Evaluación online con un electrodo frontal                                                            | Entrenamiento y test en F3-A2                                       |
| <i>ISRUC02</i>     | Evaluación online con un electrodo central                                                            | Entrenamiento y test en C3-A2                                       |
| <i>ISRUC03</i>     | Evaluación online con un electrodo occipital                                                          | Entrenamiento y test en O1-A2                                       |
| <i>ISRUC04</i>     | Evaluación online combinando varios canales EEG de la región frontal, central y occipital             | Entrenamiento y test en F3-A2 + C3-A2 + O1-A2                       |
| <i>ISRUC05</i>     | Evaluación online sin información futura combinando varios canales EEG y utilizando el EMG, EKG y EOG | Entrenamiento y test en F3-A2 + C3-A2 + O1-A2 + EMG + EKG + EOG (2) |

Tabla 3.2: Resumen de las pruebas planteadas sobre los datasets públicos.

## Capítulo 4

# Resultados y discusión

Los objetivos del presente capítulo consisten en describir y analizar el conjunto de resultados obtenido en cuanto al rendimiento del modelo desarrollado (sección 3.1) sobre los datasets disponibles (apartado 3.3.1) siguiendo la metodología de evaluación propuesta (apartado 3.3.2), correspondiente a la sección 4.1 de este capítulo, así como los resultados asociados a las técnicas de explicabilidad aplicadas (sección 3.2), expuestos en la sección 4.2.

### 4.1. Clasificación automática de las fases del sueño

La evaluación del rendimiento del modelo de aprendizaje profundo sobre los datasets públicos SleepEDF, DOD-H, DOD-O e ISRUC-Sleep, y sobre los datos recopilados por Bitbrain en los conjuntos *Study 1* y *Study 3*, se realiza en los apartados 4.1.1 y 4.1.2, respectivamente, a través de una tabla en la que se exhibe el informe de clasificación obtenido respecto a las métricas de *accuracy*, *precision*, *recall* y *F1-score* para cada prueba asociada al dataset correspondiente, junto con las matrices de confusión generadas, las cuales contienen información acerca de las predicciones realizadas en referencia a cada etapa del sueño.

#### 4.1.1. Datasets públicos

##### SleepEDF

En primer lugar, en cuanto al dataset SleepEDF, los resultados obtenidos (ver Tabla 4.1) muestran que la información del canal Fpz-Cz obtiene un rendimiento superior a la del canal Pz-Oz, con una diferencia de *accuracy* del 3% y el 4% en las formas de evaluación *offline* y *online*, respectivamente (83.09% - 80.27% y 81.29% - 77.32%). Además, para todos los tests, los valores de las métricas computadas son bastante parejos entre sí, aunque la desviación típica ( $\sigma$ ) de la *accuracy* entre sujetos se encuentra en el límite considerado como óptimo.

Por otro lado, la evaluación *online* implica una disminución del rendimiento respecto a la evaluación *offline*, con una caída de la *accuracy* del 2% en el canal Fpz-Cz y del 3% en el canal Pz-Oz. Sin embargo, teniendo en cuenta que existe una diferencia de ventana temporal de 100 épocas con información pasada y futura, frente a un contexto de 5 épocas con únicamente información temporal pasada, los resultados obtenidos son considerablemente buenos y prometedores. Además, las métricas calculadas se encuentran por encima (*SEDF00*, *SEDF01*, *SEDF02*) y a la altura (*SEDF03*) del grado de acuerdo entre evaluadores humanos definido para una misma grabación (70%-80% [21], [22]).

Adicionalmente, ya que se trata de uno de los datasets más utilizado en la literatura, es posible comparar el rendimiento obtenido frente al reportado en otros modelos existentes. En

consecuencia, el rango de valores respecto a la *accuracy* oscila entre el 80% y 87% en el canal Fpz-Cz (ver Tabla 2.1), de forma que el rendimiento arrojado por el modelo desarrollado está a la altura de los resultados de los modelos presentes en el estado del arte.

| SleepEDF      | Accuracy                    | Precision | Recall | F1-score |
|---------------|-----------------------------|-----------|--------|----------|
| <i>SEDF00</i> | 0.8309 ( $\sigma = 0.073$ ) | 0.8535    | 0.8292 | 0.8304   |
| <i>SEDF01</i> | 0.8027 ( $\sigma = 0.058$ ) | 0.8231    | 0.8027 | 0.8008   |
| <i>SEDF02</i> | 0.8129 ( $\sigma = 0.078$ ) | 0.8421    | 0.8129 | 0.8127   |
| <i>SEDF03</i> | 0.7732 ( $\sigma = 0.054$ ) | 0.7906    | 0.7732 | 0.7624   |

Tabla 4.1: Resultados del informe de clasificación sobre el dataset SleepEDF.

Por último, las matrices de confusión de la Figura 4.1 ilustran que los errores se concentran en la fase N1 del sueño, mientras que el resto de etapas, en general, se mantienen estables, siendo las fases W, N3 y REM las mejor clasificadas, seguidas de la etapa N2. También, se puede denotar que los tests sobre el canal Fpz-Cz arrojan una precisión mayor que el canal Pz-Oz, y que la evaluación *offline* es ligeramente superior a la *online*, tal y como muestra el informe de clasificación.

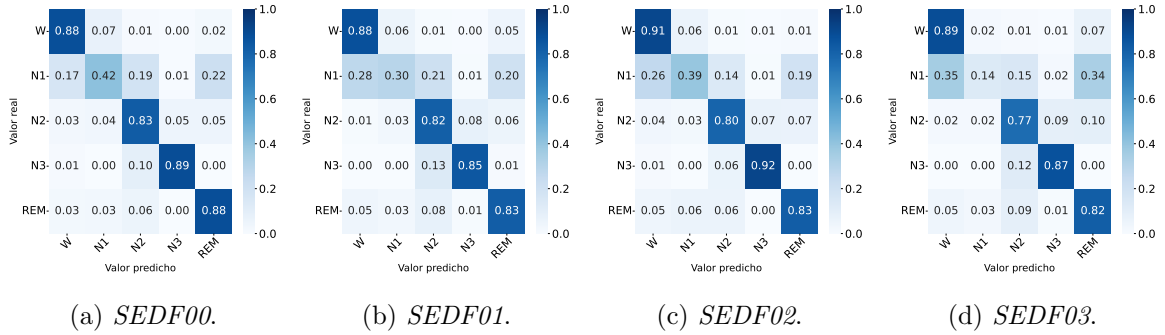


Figura 4.1: Matrices de confusión para cada test realizado sobre el dataset SleepEDF.

## DOD-H

Para empezar, el test *offline* sobre el canal F3-M2 arroja una *accuracy* del 76.89% con una variabilidad ( $\sigma$ ) elevada (ver Tabla 4.2). Las siguientes pruebas *online* sobre los canales F3-M2, C3-M2 y O1-M2, respectivamente, muestran que el rendimiento en cuanto a la métrica de *accuracy* de la información de un canal frontal (79.09%) es superior a la de las regiones central (75.91%) y occipital (65.94%). Además, se obtiene una eficiencia superior a través de la evaluación *online* del canal F3-M2 respecto a la *offline*, con una diferencia de aproximadamente un 2% de *accuracy*, lo cual se puede deber a la elevada desviación típica existente entre los diferentes *folds* del test *DODH00*, ya que lo esperable y lo que ocurre en el resto de pruebas del trabajo es que la evaluación en tiempo real resulte inferior a la *offline*, tratándose de un fenómeno relativamente aislado.

Por otro lado, en el test *DODH04* se combinan los canales F3-M2, C3-M2 y O1-M2, lo cual no produce un aumento en la precisión respecto a utilizar únicamente el canal frontal F3-M2

(-3%), de forma que la información de una única fuente puede ser suficiente, aunque sí que se mantienen los resultados con el canal central C3-M2 y se incrementan en referencia al canal occipital O1-M2 (+10%). Sin embargo, combinar estos canales EEG junto con la información del EMG, EKG y EOG provoca una caída significativa en el rendimiento obtenido (-10% respecto al test *DODH04*). A excepción de las pruebas *DODH03* y *DODH05*, todos los resultados se encuentran en el rango de acuerdo entre evaluadores humanos (70%-80%).

| DOD-H         | Accuracy                    | Precision | Recall | F1-score |
|---------------|-----------------------------|-----------|--------|----------|
| <i>DODH00</i> | 0.7689 ( $\sigma = 0.111$ ) | 0.8256    | 0.7689 | 0.7704   |
| <i>DODH01</i> | 0.7909 ( $\sigma = 0.064$ ) | 0.8213    | 0.7908 | 0.7770   |
| <i>DODH02</i> | 0.7591 ( $\sigma = 0.076$ ) | 0.7985    | 0.7591 | 0.7501   |
| <i>DODH03</i> | 0.6594 ( $\sigma = 0.072$ ) | 0.7471    | 0.6593 | 0.6538   |
| <i>DODH04</i> | 0.7586 ( $\sigma = 0.094$ ) | 0.8174    | 0.7586 | 0.7555   |
| <i>DODH05</i> | 0.6621 ( $\sigma = 0.103$ ) | 0.7359    | 0.6621 | 0.6592   |

Tabla 4.2: Resultados del informe de clasificación sobre el dataset DOD-H.

Finalmente, en cuanto a las matrices de confusión de la Figura 4.2, se puede apreciar que la peor fase del sueño en todos los tests es la N1. Además, en las pruebas *DODH00*, *DODH01*, *DODH02* y *DODH04*, las etapas W, N3 y REM se mantienen bastante estables, con buenos resultados, seguidas de la etapa N2, la cual presenta un nivel ligeramente superior de confusiones. Sin embargo, en los tests *DODH03* y *DODH05*, prácticamente todas las fases experimentan una cantidad de confusiones mayor respecto al resto de pruebas, mostrando valores inferiores en cuanto a la precisión obtenida.

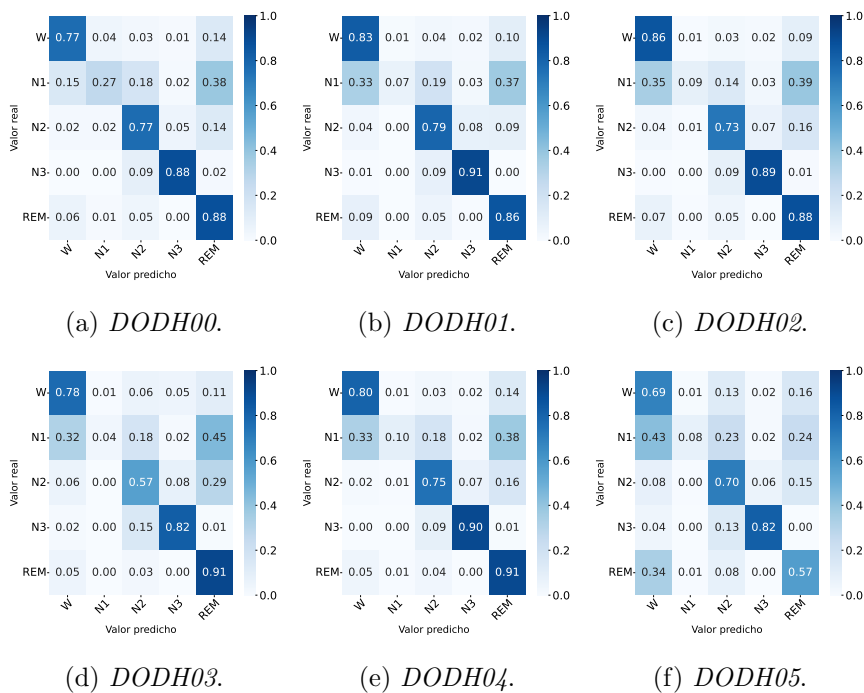


Figura 4.2: Matrices de confusión para cada test realizado sobre el dataset DOD-H.

## DOD-O

En primer lugar, en la Tabla 4.3 se puede apreciar que la primera prueba *offline* sobre el canal F3-M2 obtiene el mejor resultado, como es de esperar en una evaluación *offline*, arrojando una *accuracy* del 78.01%. En siguiente lugar, los tests *online* sobre los canales F3-M2, C3-M2 y O1-M2, respectivamente, exhiben que la *accuracy* de las regiones frontal (75.07%) y central (74.59%) es superior a la experimentada en la zona occipital (60.48%), sufriendo una caída del 15%. Por consiguiente, en el mejor de los casos, la evaluación *online* supone un decremento del 3% de la *accuracy* respecto a la evaluación *offline*, lo cual se puede considerar un buen resultado.

Por otra parte, combinar la información de los canales F3-M2, C3-M2 y O1-M2 (test *DODO04*) mantiene los resultados obtenidos utilizando únicamente el canal F3-M2 o el canal C3-M2, aunque incrementa el rendimiento respecto al canal O1-M2 (+15%). Además, añadir información sobre el EMG, EKG y EOG, junto con estos canales, no aumenta la precisión del modelo en la clasificación, sino que se obtienen prácticamente los mismos resultados (74.36%). Cabe destacar que todos las pruebas, a excepción del test *DODO03*, se encuentran en el rango de consenso entre evaluadores humanos (70%-80%).

| DOD-O         | Accuracy                    | Precision | Recall | F1-score |
|---------------|-----------------------------|-----------|--------|----------|
| <i>DODO00</i> | 0.7801 ( $\sigma = 0.067$ ) | 0.8022    | 0.7801 | 0.7802   |
| <i>DODO01</i> | 0.7507 ( $\sigma = 0.051$ ) | 0.7843    | 0.7507 | 0.7438   |
| <i>DODO02</i> | 0.7459 ( $\sigma = 0.060$ ) | 0.7815    | 0.7459 | 0.7368   |
| <i>DODO03</i> | 0.6048 ( $\sigma = 0.101$ ) | 0.6605    | 0.6048 | 0.5829   |
| <i>DODO04</i> | 0.7533 ( $\sigma = 0.061$ ) | 0.7932    | 0.7533 | 0.7469   |
| <i>DODO05</i> | 0.7436 ( $\sigma = 0.070$ ) | 0.7619    | 0.7436 | 0.7315   |

Tabla 4.3: Resultados del informe de clasificación sobre el dataset DOD-O.

Por último, respecto a las matrices de confusión de la Figura 4.3, la fase N1 vuelve a ser, con diferencia, la peor clasificada en todas las pruebas. El resto de etapas obtienen resultados considerablemente buenos, donde destaca la precisión sobre las fases W, N3 y REM, seguidas de N2. Sin embargo, el test *DODO03* es el único que arroja un nivel alto de confusiones en todas las fases, a excepción de la etapa W, constituyendo la prueba con el rendimiento más bajo obtenido sobre este conjunto de datos.

## ISRUC-Sleep

La descripción y el análisis del informe de clasificación y las matrices de confusión generadas a partir de la ejecución del modelo de aprendizaje profundo sobre el dataset ISRUC-Sleep se ilustran en el Apéndice E.

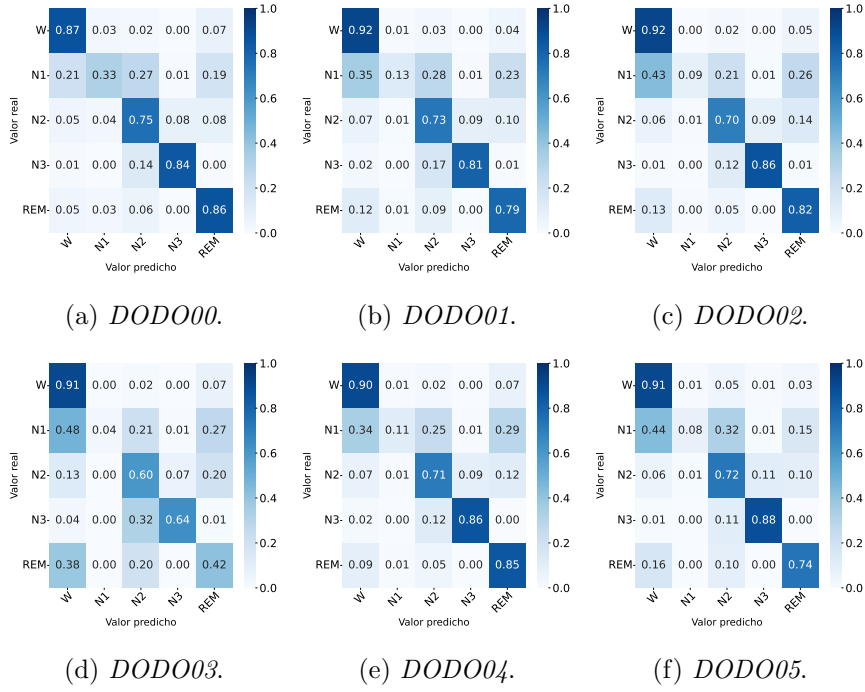


Figura 4.3: Matrices de confusión para cada test realizado sobre el dataset DOD-O.

## Discusión general

Una vez analizados los resultados obtenidos en cada dataset público, resulta de relevancia destacar los siguientes aspectos. En primer lugar, se ha determinado que el rendimiento del modelo desarrollado está a la altura de los resultados reportados en la literatura, garantizando la efectividad del sistema en la clasificación automática del sueño. Respecto a estos resultados, la etapa con más errores en su clasificación ha sido la fase N1 del sueño, confundida ampliamente con las etapas W y N2, ya que se considera una fase de transición entre ambas y puede ser etiquetada fácilmente de forma incorrecta, teniendo en cuenta que constituye únicamente el 5% de la noche. En cambio, las etapas W, N3 y REM tienden a ser las más sólidas en cuanto a resultados, aunque también son propensas a sufrir confusiones en su evaluación.

Por otro lado, se ha determinado que la posición de los electrodos es de gran importancia, destacando los resultados de las regiones prefrontal (Fp), frontal (F), central (C) y parietal (P) frente a la zona occipital (O). Además, el uso de varios canales y su combinación con información sobre el EMG, EKG y EOG no supone un impacto considerable en el rendimiento, manteniendo o decrementando la precisión arrojada, de forma que una única fuente de información puede ser suficiente para realizar la clasificación del sueño.

Adicionalmente, la evaluación en línea no implica una caída significativa en el rendimiento respecto a una metodología *offline*, obteniendo unos resultados prometedores para los objetivos planteados en este proyecto. De hecho, una configuración *online* basada en un único canal EEG, que constituye el escenario más limitante en cuanto a la cantidad de información disponible, puede ser suficiente para realizar la clasificación automática del sueño, favoreciendo la simplicidad y escalabilidad de la tecnología resultante, así como la aceleración de su adopción en entornos clínicos y domésticos.

Finalmente, cabe mencionar que el rendimiento obtenido en los datasets DOD-H y DOD-O es bastante similar, demostrando que el modelo es capaz de clasificar eficazmente el sueño en sujetos con trastornos respiratorios relacionados con el sueño. Sin embargo, en el conjunto ISRUC-Sleep (ver Apéndice E) el rendimiento obtenido es inferior al experimentado en el dataset DOD-H, produciendo que el modelo no sea capaz de obtener una *accuracy* dentro del rango de acuerdo entre evaluadores humanos para los sujetos no sanos de este subconjunto. De esta forma, se ha cumplido el objetivo de evaluar los resultados arrojados en sujetos con patologías del sueño.

#### 4.1.2. Datos recopilados por Bitbrain

##### Bitbrain · Study 1

En primer lugar, los resultados obtenidos sobre el conjunto *Study 1* de Bitbrain (ver Tabla 4.4) muestran que los datos etiquetados a partir del consenso de 3 expertos produce un rendimiento significativamente superior al conjunto evaluado por un único técnico, con una diferencia de *accuracy* del 7% y el 5% en las formas de evaluación *offline* y *online*, respectivamente (80.29% - 87.78% y 79.51% - 84.31%). En consecuencia, se ha determinado que el aumento de profesionales que puntúan las grabaciones conlleva un incremento en la precisión del modelo, debido a que el acuerdo entre diferentes expertos implica un nivel de objetividad mayor, y por lo tanto una clasificación de calidad superior.

Por otra parte, la evaluación en tiempo real produce una ligera disminución del rendimiento respecto a la metodología *offline*, con una caída de *accuracy* del 1% para el conjunto evaluado por un único experto y del 3% sobre el grupo etiquetado a partir del consenso de 3 técnicos. Por consiguiente, los resultados son considerablemente buenos, ya que la diferencia entre los valores es mínima. Además, las métricas computadas se encuentran por encima del rango de acuerdo entre evaluadores humanos (70%-80%), enfatizando la efectividad alcanzada por el modelo. También, destaca la *accuracy* obtenida con la prueba *STD101* (87.78%), ya que se constituye como el valor más alto arrojado en todos los tests ejecutados.

| Bitbrain · Study 1 | Accuracy                    | Precision | Recall | F1-score |
|--------------------|-----------------------------|-----------|--------|----------|
| <i>STD100</i>      | 0.8029 ( $\sigma = 0.043$ ) | 0.8230    | 0.8029 | 0.8029   |
| <i>STD101</i>      | 0.8778 ( $\sigma = 0.014$ ) | 0.8844    | 0.8778 | 0.8773   |
| <i>STD102</i>      | 0.7951 ( $\sigma = 0.034$ ) | 0.8136    | 0.7951 | 0.7876   |
| <i>STD103</i>      | 0.8431 ( $\sigma = 0.033$ ) | 0.8557    | 0.8431 | 0.8372   |

Tabla 4.4: Resultados del informe de clasificación sobre el dataset Study 1 de Bitbrain.

Por último, en la Figura 4.4 se ilustran las matrices de confusión generadas, en las que, de nuevo, la fase N1 es la que más confusiones genera, siendo clasificada incorrectamente como las etapas W, N2 y REM. En general, debido al buen rendimiento obtenido a través de las pruebas planteadas en este dataset, se puede afirmar que todas las matrices son óptimas para el resto de fases del sueño (W, N2, N3 y REM).

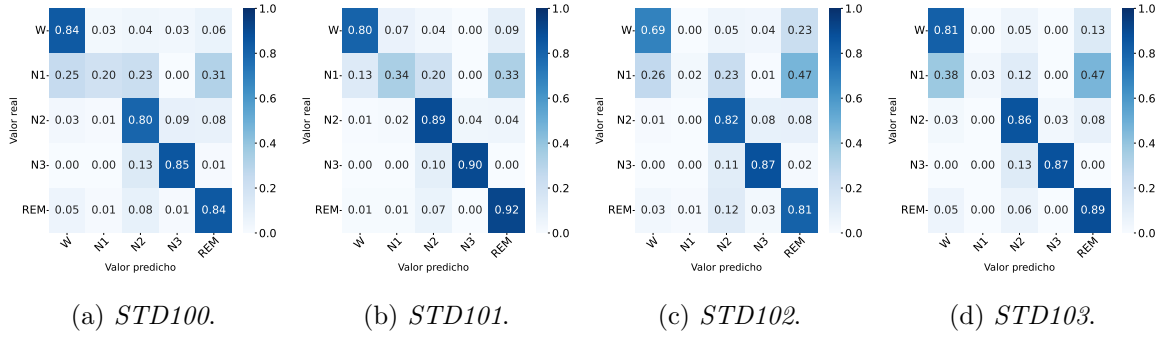


Figura 4.4: Matrices de confusión para cada test realizado sobre el dataset Study 1 de Bitbrain.

### Bitbrain · Study 3

Los resultados obtenidos en el *Study 3* de Bitbrain, ilustrados en la Tabla 4.5, muestran que el rendimiento arrojado en cuanto a *accuracy* en los canales AF7 + AF8 del EEG de la banda textil (77.85% y 77.14%) son ligeramente superiores a los del PSG en los canales F3 + F4 (77.33% y 75.02%), tanto en la configuración *offline* como en línea. Por consiguiente, debido a que el PSG es el dispositivo de referencia utilizado en la medicina del sueño, es posible afirmar que la calidad de la señal de la diadema textil se encuentra a la altura, determinando que esta nueva tecnología es capaz de medir la actividad cerebral con una precisión más que suficiente para llevar a cabo la clasificación automática de las etapas del sueño.

Además, la aplicación de una metodología *online* sobre los canales EEG de la banda no tiene un impacto negativo sobre el rendimiento, manteniendo prácticamente los mismos resultados respecto a una evaluación *offline*. En consecuencia, este experimento demuestra la efectividad que supondría la integración del modelo de aprendizaje profundo con la tecnología textil inteligente desarrollada, con la posibilidad de evaluar las fases del sueño en tiempo real en entornos domésticos. En cambio, para los datos del PSG se experimenta una caída de la *accuracy* del 2%, lo cual se puede considerar también un resultado óptimo. Adicionalmente, las métricas computadas presentan valores muy similares entre sí, junto con una desviación típica ( $\sigma$ ) bastante baja, y todos los resultados se encuentran en el rango de consenso entre evaluadores humanos (70%-80%).

| Bitbrain · Study 3 | Accuracy                    | Precision | Recall | F1-score |
|--------------------|-----------------------------|-----------|--------|----------|
| <i>STD300</i>      | 0.7733 ( $\sigma = 0.019$ ) | 0.7809    | 0.7733 | 0.7658   |
| <i>STD301</i>      | 0.7785 ( $\sigma = 0.026$ ) | 0.7865    | 0.7785 | 0.7719   |
| <i>STD302</i>      | 0.7502 ( $\sigma = 0.051$ ) | 0.7650    | 0.7502 | 0.7396   |
| <i>STD303</i>      | 0.7714 ( $\sigma = 0.031$ ) | 0.7794    | 0.7714 | 0.7591   |

Tabla 4.5: Resultados del informe de clasificación sobre el dataset Study 3 de Bitbrain.

Finalmente, la fase N1 del sueño se constituye como la peor clasificada, como se puede observar en la Figura 4.5, siendo ampliamente confundida con las etapas W, N2 y REM. Por otro lado, las fases N3 y REM experimentan una precisión superior, aunque son puntuadas incorrectamente como la etapa N2 en muchas ocasiones, mientras que las fases W y N2 son las que mejores resultados arrojan respecto al resto.

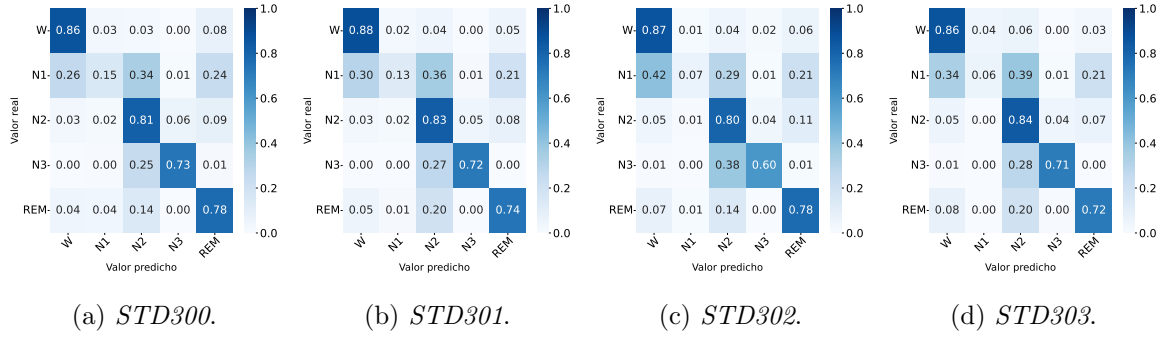


Figura 4.5: Matrices de confusión para cada test realizado sobre el dataset Study 3 de Bitbrain.

## 4.2. Explicabilidad e interpretabilidad

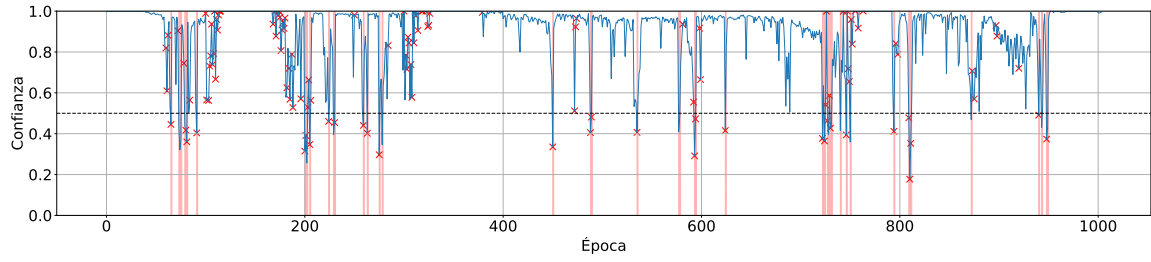
Una vez realizados la evaluación y el análisis del rendimiento arrojado por el modelo de aprendizaje profundo en diferentes conjuntos de datos, se van a ilustrar los resultados obtenidos mediante la aplicación de las técnicas de explicabilidad e interpretabilidad propuestas en la sección 3.2 del capítulo anterior. Para ello, se va a utilizar la grabación del sujeto 14 (SC414) del dataset SleepEDF, utilizando el canal Fpz-Cz del EEG, sobre el que se obtiene una *accuracy* del 88.35% siguiendo una metodología de evaluación *offline*. Además, por completitud, en el Apéndice F se muestran resultados adicionales, correspondientes al sujeto 5 (SC405) del mismo dataset, utilizando el canal Pz-Oz del EEG.

### Cuantificación de la confianza

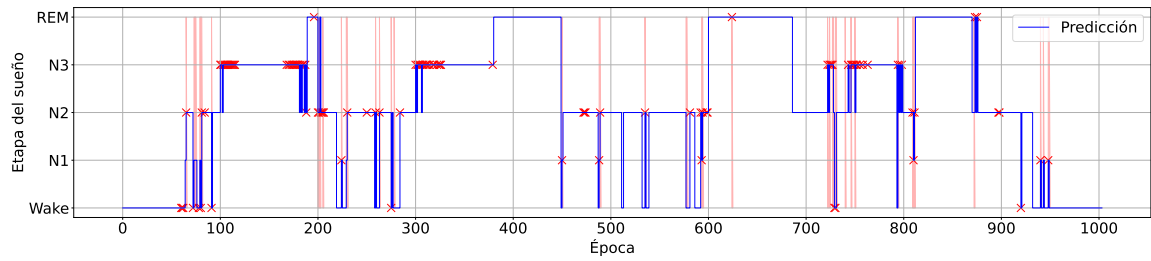
En primer lugar, en la Figura 4.6 se ilustran los resultados obtenidos para el estudio del nivel de confianza del modelo en la clasificación del sueño, con un umbral (*threshold*) definido con el valor 0.5. En consecuencia, en la Figura 4.6a, donde se muestran los valores de confianza para cada época calculados a partir de las ecuaciones 3.3 y 3.4, se puede apreciar que los fallos se concentran en zonas en las que el modelo arroja una confianza baja, de forma que muchas de las épocas que experimentan una incertidumbre por encima del umbral establecido son clasificadas incorrectamente (concretamente el 65%).

Adicionalmente, existe la posibilidad de que las épocas clasificadas con un nivel de confianza por debajo del *threshold* definido sean evaluadas de forma manual por profesionales. Por consiguiente, el valor del umbral puede establecerse en base a las necesidades del estudio correspondiente, adaptándose a la tolerancia a fallos permitida y a la posibilidad de recurrir a expertos en épocas que experimentan un nivel de incertidumbre elevado.

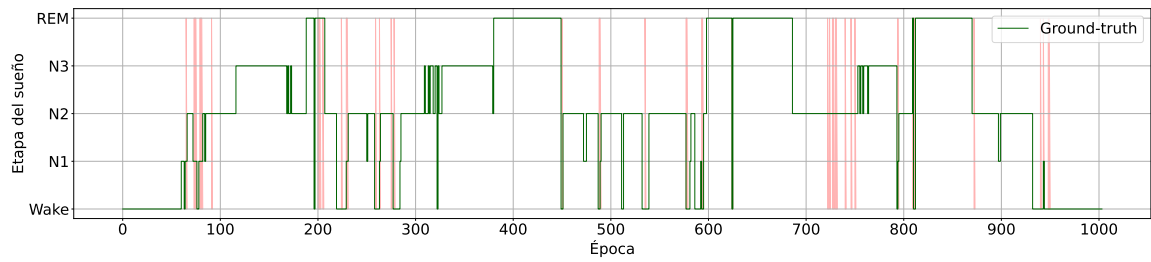
Por otro lado, en la Figura 4.6b se exhibe el hipnograma asociado a las predicciones del algoritmo, en el que se puede observar que los fallos se concentran en las zonas en las que el nivel de confianza está por debajo del umbral definido (sombreadas en rojo) y en las regiones de transición entre etapas del sueño. Además, cabe mencionar que el bajo número de estados N1 se clasifica incorrectamente con frecuencia, mientras que muchas de las etapas N2 se confunden en varias ocasiones con la fase N3. Respecto a la Figura 4.6d, que muestra el hipnograma de las predicciones y el hipnograma del Ground-truth (Figura 4.6c), se ha determinado que el solape entre ellos es considerablemente bueno, generando una evaluación de la grabación muy similar e indicando el buen rendimiento arrojado sobre el sujeto analizado.



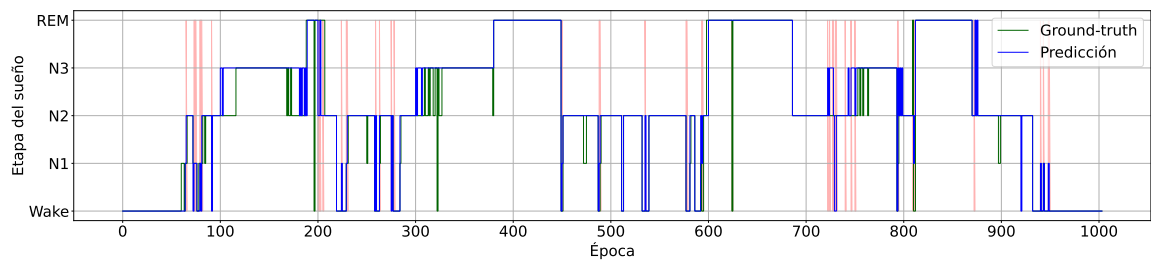
(a) Cuantificación de la confianza, donde  $\times$  representa los fallos ( $threshold = 0.5$ ).



(b) Hipnograma asociado a las predicciones del modelo, donde  $\times$  representa los fallos.



(c) Hipnograma asociado al Ground-truth.



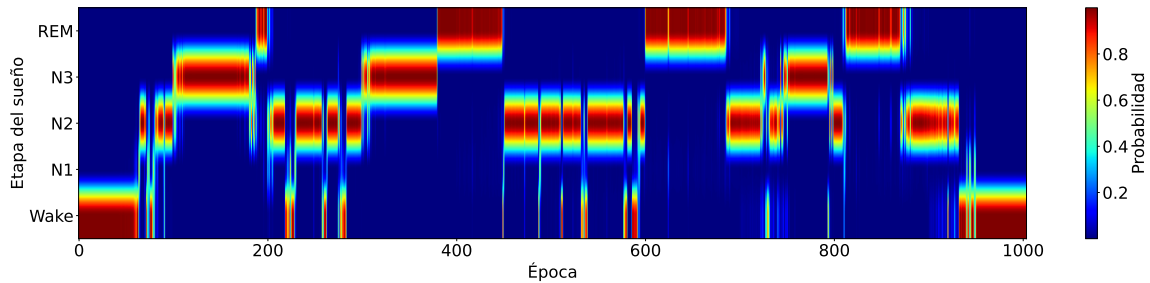
(d) Hipnograma correspondiente a las predicciones del modelo y el Ground-truth.

Figura 4.6: Ilustración de la estimación de la confianza del modelo en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG ( $accuracy = 88.35\%$ , evaluación *offline*).

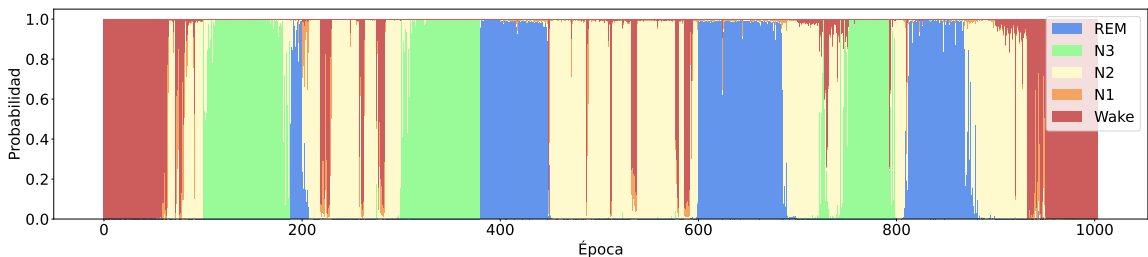
## Análisis probabilístico

En lo que se refiere al análisis de las probabilidades de salida del modelo, en la Figura 4.7a se ilustra, en primer lugar, el mapa de calor obtenido a partir de la intensidad de los valores arrojados por el algoritmo. Uno de los aspectos más destacables que se obtiene a partir de este gráfico se basa en la semejanza que presenta respecto al hipnograma generado (ver Figura 4.6b), determinando la relación existente entre las probabilidades de salida y la clasificación realizada. Además, se puede apreciar que los fallos se concentran en las regiones con un elevado número de discontinuidades, que indican una distribución más uniforme de la probabilidad y un consecuente nivel de confianza inferior en la predicción, junto con las zonas de transición entre fases del sueño, de la misma forma que se experimentaba al analizar la incertidumbre del modelo.

Por otra parte, la Figura 4.7b muestra un gráfico de barras en el que se observa la distribución de las probabilidades de cada fase del sueño asociadas a las épocas de la grabación. En consecuencia, las regiones que concentran colores únicos se asocian con una distribución de valores centrada en una clase, lo cual implica un nivel de confianza elevado en la predicción, como se ha visto en el apartado anterior, y un consecuente rendimiento superior. En cambio, las zonas cuyos colores tienden a mezclarse, normalmente en regiones de transición entre etapas del sueño, experimentan una distribución de valores más uniforme, indicando un nivel de incertidumbre superior y una consecuente efectividad inferior. De esta manera, se concluye que el análisis probabilístico está directamente relacionado con la cuantificación de la confianza del modelo, obteniendo resultados comunes respecto al entendimiento de la salida del modelo de aprendizaje profundo en la clasificación del sueño.



(a) Mapa de calor asociado a cada época de la grabación en el que se ilustra la intensidad de los valores de probabilidad de las etapas del sueño.



(b) Gráfico de barras correspondiente a cada época de la grabación en el que se muestra un vector de colores según la intensidad de las probabilidades de las fases del sueño.

Figura 4.7: Análisis de las probabilidades de salida del modelo en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG (*accuracy* = 88.35%, evaluación *offline*).

## Grad-CAM

El objetivo de la aplicación del método Grad-CAM sobre el modelo consiste en relacionar los patrones de activación de las capas convolucionales de la red con las reglas visuales definidas en el manual de la AASM [33] para cada fase del sueño, detalladas en la sección 1.4 y resumidas en la Tabla 4.6. De esta forma, se aporta explicabilidad en la codificación a nivel de época, correspondiente al componente CNN de la arquitectura desarrollada.

| Etapa | Reglas                                                                                         |
|-------|------------------------------------------------------------------------------------------------|
| Wake  | Presencia de más del 50 % de frecuencias alpha (8-12 Hz).                                      |
| N1    | Existencia de ondas theta (4-8 Hz) de amplitud baja.                                           |
| N2    | Aparición de uno o más complejos K y husos del sueño.                                          |
| N3    | Predominio de ondas delta (1-4 Hz), con la posible aparición de husos del sueño y complejos K. |
| REM   | Manifiesta ondas en diente de sierra, con un EEG de amplitud baja y frecuencia mixta.          |

Tabla 4.6: Resumen de las reglas visuales del manual de la AASM.

Para ello, en la Figura 4.8 se ilustra una época representativa de la grabación para cada fase del sueño, identificando las regiones (características) de la señal que son más relevantes para el modelo en relación con las indicaciones de la AASM. Por consiguiente, en primer lugar, para la fase Wake, se puede observar que la activación se concentra en todas las zonas de la época, reflejando que la activación se centra en identificar las ondas alfa (8-12 Hz), las cuales caracterizan esta etapa.

A continuación, en la fase N1 también se puede observar que la activación se distribuye en los diferentes puntos temporales de la señal, reflejando que el modelo da importancia a las ondas theta (4-8 Hz) correspondientes. Además, en estas figuras se puede apreciar la similitud que presentan las etapas Wake y N1, siendo esta última considerada como una transición entre la vigilia y el sueño, lo cual sumado a que se constituye como la clase con la menor cantidad de muestras (5%), provoca el elevado número de confusiones que se han experimentado en las pruebas ejecutadas en la sección 4.1.

Por otro lado, en la fase N2 se puede apreciar una relación evidente entre la activación de la capa convolucional y las reglas de la AASM, donde se exhiben los denominados husos del sueño (*sleep spindles*) y eventos complejos K (*K-complex*). En siguiente lugar, en la etapa N3 predominan las ondas delta (1-4 Hz), distribuidas en numerosas regiones de la señal. Finalmente, el modelo identifica las ondas en diente de sierra (*sawtooth waves*) presentes en la fase REM del sueño.

En conclusión, se ha desarrollado un método que permite identificar los patrones de activación aprendidos por la red en el componente de codificación a nivel de época, con el propósito de relacionarlos con las reglas del manual de la AASM. Adicionalmente, esta técnica posibilita que los expertos validen si el modelo examina las mismas características del EEG que ellos buscan al evaluar las épocas de una grabación, incrementando la fiabilidad de estos sistemas e impulsando su aceptación y adopción en el sector médico.

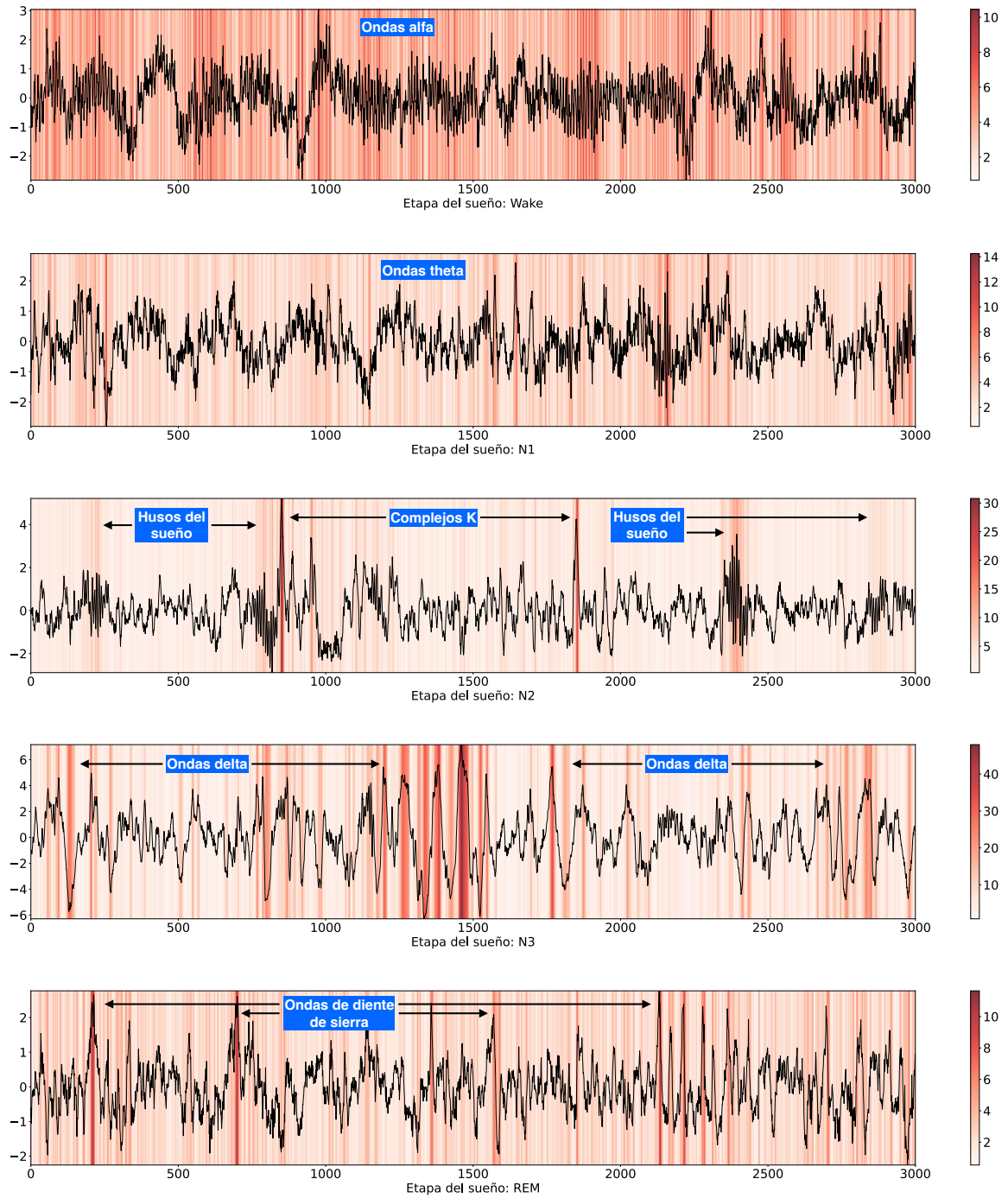


Figura 4.8: Resultados obtenidos mediante la aplicación del método Grad-CAM sobre el modelo (segunda capa convolucional de la red), indicando las regiones de las épocas que más activación experimentan en relación con las reglas definidas en el manual de la AASM, correspondientes a cada fase del sueño. La ejecución se ha realizado en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG ( $accuracy = 88.35\%$ , evaluación *offline*).

## Influencia de las épocas vecinas

Con el propósito de determinar las épocas vecinas de la secuencia de entrada a las que presta mayor atención el modelo cuando clasifica la época objetivo, se han utilizado los pesos de autoatención generados en las capas Multi-Head Self-Attention, pertenecientes al componente Transformer de la red, aportando interpretabilidad a nivel de codificación secuencial. De esta manera, en la Figura 4.9 se ilustran los resultados obtenidos para algunas de las épocas de la primera secuencia de la grabación ( $L = 100$  épocas).

En primer lugar, se puede apreciar que los pesos de autoatención de las primeras épocas se concentran en la primera mitad de la secuencia de entrada, aunque también hacen uso, en menor medida, de las épocas contenidas en la segunda mitad. Por otro lado, las épocas más centrales de la secuencia atienden a información tanto pasada como futura, concentrando una cantidad mayor de los valores de autoatención en la segunda mitad. Este aspecto fundamenta las diferencias existentes en el rendimiento de la metodología *offline* y en tiempo real, ya que la información futura resulta de relevancia para el contexto temporal utilizado en el proceso de clasificación. En cambio, las últimas épocas concentran la atención en la segunda mitad de la secuencia, sin hacer especial uso de la información de las primeras épocas.

Adicionalmente, la distribución de los pesos de autoatención no solo tiene en cuenta la información posicional o contextual de la secuencia de entrada, sino que también se puede observar que, en cierta manera, captura las transiciones entre las fases del sueño. En consecuencia, los valores de autoatención cambian significativamente cuando el sujeto analizado comienza el periodo del sueño, de forma que en las épocas 50 y 60 (Wake) el modelo tiende a fijarse, en mayor medida, en información próxima (futuras etapas N1 y N2, ver hipnograma de la Figura 4.6c), a diferencia de las épocas de vigilia anteriores. Además, la asignación de pesos de autoatención en cuanto al contexto temporal puede experimentar ligeras diferencias según la fase del sueño de la que se trate, como ocurre en la época 80.

En definitiva, se trata de una técnica que permite entender el comportamiento del modelo en cuanto a la importancia que tiene el contexto temporal y la información posicional para la clasificación de una época respecto a las épocas vecinas, aportando explicabilidad a nivel de codificación secuencial (red Transformer). Por otra parte, los métodos anteriores se centran en la interpretabilidad a nivel de codificación de época (red CNN) y en el entendimiento de los valores de salida del modelo (probabilidad y confianza), constituyendo a través de todas las metodologías propuestas una visión global acerca del funcionamiento de todos los componentes del sistema desarrollado para la clasificación del sueño.

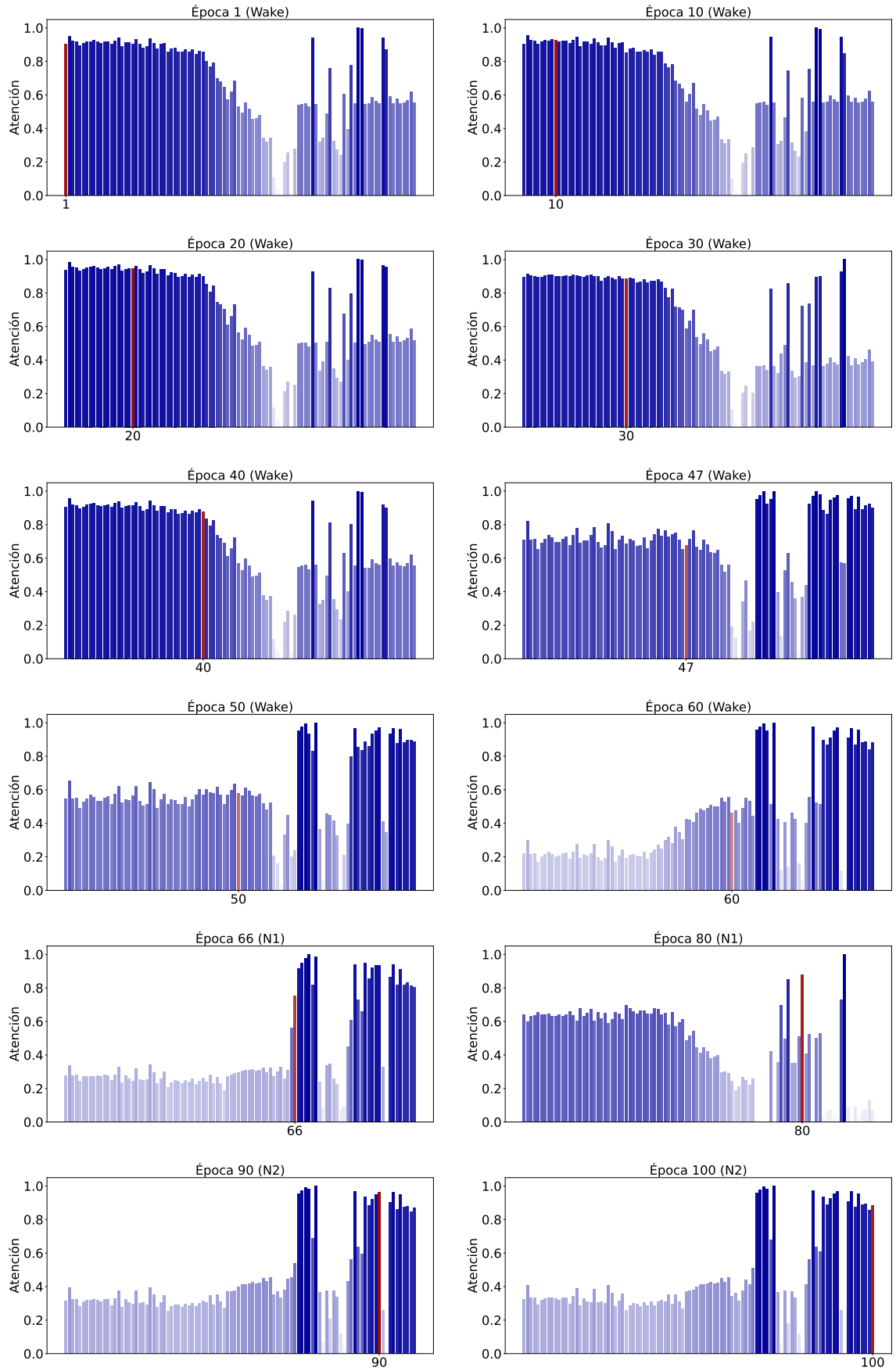


Figura 4.9: Resultados del análisis de la influencia de las épocas vecinas respecto a la clasificación de la época objetivo (segunda capa Multi-Head Self-Attention de la red) en el sujeto 14 del dataset SleepEDF utilizando el canal Fpz-Cz del EEG ( $accuracy = 88.35\%$ , evaluación *offline*).

## Capítulo 5

# Conclusiones

En el presente proyecto se ha desarrollado con rotundo éxito un modelo de aprendizaje profundo para la clasificación de las fases del sueño. Los resultados obtenidos mediante la técnica presentada, como se ha podido comprobar a partir de las diferentes pruebas realizadas y el correspondiente análisis, se encuentran a la altura del rendimiento de otras arquitecturas existentes en el estado del arte. Además, estos resultados tienen una precisión dentro del rango de valores de consenso entre evaluadores profesionales.

Adicionalmente, a diferencia de los tests planteados en los modelos de la literatura científica, se ha comprobado el rendimiento del algoritmo frente a una metodología de evaluación del sueño en tiempo real, es decir, en ausencia de información futura. De esta forma, los resultados obtenidos son considerablemente buenos, ya que no suponen una caída significativa del rendimiento respecto a una evaluación *offline* (con información pasada y futura) y se encuentran en el rango de acuerdo entre evaluadores humanos. De hecho, una configuración *online* con un único canal EEG, que constituye el escenario más limitante en cuanto a la cantidad de información disponible, puede ser suficiente para realizar la clasificación del sueño.

Por otra parte, cabe mencionar el rendimiento obtenido en los datos recopilados por Bitbrain. En primer lugar, el conjunto *Study 1* arroja la *accuracy* más alta experimentada en todo el trabajo, tanto en la forma de evaluación *offline* (87.78%) como *online* (84.31%). En siguiente lugar, el dataset *Study 3* destaca por la precisión conseguida en los datos pertenecientes a la banda textil de Bitbrain. Concretamente, para una metodología *offline*, se observa un valor de *accuracy* del 77.85%, mientras que en la evaluación en tiempo real se obtiene un 77.14%, comprendiendo valores prácticamente idénticos, sin una caída del rendimiento entre formas de evaluación y dentro del rango de consenso entre evaluadores. Por lo tanto, la calidad de la señal de la diadema textil se encuentra a la altura del PSG (dispositivo de referencia), determinando que esta nueva tecnología es capaz de medir la actividad cerebral con una precisión más que suficiente para llevar a cabo la clasificación automática de las etapas del sueño en conjunto con el modelo desarrollado en el presente proyecto.

Las técnicas de explicabilidad aplicadas sobre la arquitectura del modelo posibilitan un mayor entendimiento acerca del funcionamiento del sistema, y facilitan la aceptación y adopción de la inteligencia artificial en el sector clínico. Sin embargo, la mayoría de modelos existentes en la literatura no implementa metodologías de interpretabilidad, de manera que los enfoques presentados en este trabajo destacan en este aspecto. Por consiguiente, mediante los métodos propuestos se construye una visión global respecto al comportamiento de todos los componentes de la arquitectura del sistema: a nivel de codificación de época en la red convolucional (Grad-CAM), a nivel secuencial en el bloque Transformer (influencia de las

épocas vecinas en la clasificación de la época objetivo) y sobre la salida generada (nivel de confianza y análisis probabilístico).

En definitiva, el modelo de *deep learning* implementado obtiene resultados óptimos en la tarea de clasificación de las fases del sueño y es capaz de evaluar el sueño en tiempo real, posibilitando su integración con el nuevo dispositivo textil desarrollado por Bitbrain para la monitorización del sueño mediante la medición del EEG. En consecuencia, la sobrecarga de trabajo que sufren muchos médicos, los diagnósticos deficientes de los pacientes y las pruebas innecesarias pueden abordarse mediante el trabajo realizado en este proyecto, vinculándolo a las necesidades biomédicas del mundo real y poniendo de relieve su inherente aplicación práctica.

## 5.1. Trabajo futuro

En base a los prometedores resultados obtenidos a través de la metodología propuesta, las líneas de investigación para futuros trabajos son múltiples. En primer lugar, existe la posibilidad de trabajar en la mejora de la precisión en la clasificación de la fase N1 del sueño, ya que el número de confusiones experimentado en esta etapa es elevado. Para ello, se podrían añadir más sensores a nivel de fuentes de información, como la pulsioximetría, además de aplicar técnicas de *hard sampling* y *data augmentation* sobre las muestras de esta clase.

Por otro lado, resultaría de relevancia experimentar con otros mecanismos de codificación posicional (*positional encoding*) de la secuencia de entrada de una red Transformer, como la codificación posicional fija, trigonométrica o adaptativa. En relación con este aspecto, también sería importante investigar sobre la viabilidad de la aplicación de máscaras de atención sobre el componente Transformer, como las máscaras de atención causal, las cuales aseguran que los elementos de la secuencia de entrada tengan únicamente acceso a información de posiciones anteriores, garantizando la ausencia de información futura durante el entrenamiento del modelo en el bloque Transformer.

Respecto al tema de la explicabilidad, se ha realizado un estudio explorativo de las técnicas presentadas en varios sujetos específicos, de manera que existe la posibilidad de trabajar para conseguir un análisis más general y automatizado, así como estudiar la efectividad de la forma en la que se presenta la información obtenida. Adicionalmente, es necesario investigar acerca de la cuantificación de la interpretabilidad de un sistema de inteligencia artificial, determinando el límite del entendimiento humano respecto a la comprensión del funcionamiento de un modelo de aprendizaje profundo como el propuesto en este trabajo. De este modo, sería posible esclarecer hasta qué punto es necesario destapar el comportamiento de caja negra de este tipo de algoritmos para impulsar su aceptación y adopción en el ámbito sanitario.

Finalmente, se plantea la integración del modelo desarrollado en la plataforma de software de Bitbrain, de manera que se evalúe el algoritmo con los datos recopilados en tiempo real por el dispositivo textil inteligente en el entorno práctico final, conformando el sistema completo de monitorización del sueño correspondiente. Una vez realizada esta integración, el sistema podría utilizarse para llevar a cabo intervenciones en bucle cerrado para diversos tipos de aplicaciones, como pueden ser la rehabilitación de funciones perdidas o la mejora de las capacidades cognitivas. Asimismo, siguiendo esta línea, sería interesante investigar el rendimiento arrojado por el modelo en los datos adquiridos a través de la banda textil de Bitbrain sobre sujetos con patologías del sueño, con el propósito de garantizar unos resulta-

dos al nivel de los obtenidos en personas sanas, así como estimar la cantidad de información necesaria en el entrenamiento del modelo para asegurar o superar una calidad de evaluación a la altura de un experto.

# Bibliografía

- [1] J. M. Siegel, «Clues to the functions of mammalian sleep,» *Nature*, vol. 437, n.º 7063, págs. 1264-1271, 2005.
- [2] P. Maquet, «The Role of Sleep in Learning and Memory,» *Science*, vol. 294, n.º 5544, págs. 1048-1052, 2001.
- [3] L. Besedovsky, T. Lange y M. Haack, «The Sleep-Immune Crosstalk in Health and Disease,» *Physiological Reviews*, vol. 99, n.º 3, págs. 1325-1380, 2019.
- [4] F. P. Cappuccio y M. A. Miller, «Sleep and Cardio-Metabolic Disease,» *Current Cardiology Reports*, vol. 19, n.º 11, 2017.
- [5] E. C. Harding, N. P. Franks y W. Wisden, «Sleep and thermoregulation,» *Current Opinion in Physiology*, vol. 15, págs. 7-13, 2020.
- [6] S. Diekelmann y J. Born, «The memory function of sleep,» *Nature Reviews Neuroscience*, vol. 11, n.º 2, págs. 114-126, 2010.
- [7] R. Stickgold, «Sleep-dependent memory consolidation,» *Nature*, vol. 437, n.º 7063, págs. 1272-1278, 2005.
- [8] E. B. Simon, R. Vallat, C. M. Barnes y M. P. Walker, «Sleep Loss and the Socio-Emotional Brain,» *Trends in Cognitive Sciences*, vol. 24, n.º 6, págs. 435-450, 2020.
- [9] M. P. Walker, «The Role of Sleep in Cognition and Emotion,» *Annals of the New York Academy of Sciences*, vol. 1156, n.º 1, págs. 168-197, 2009.
- [10] F. P. Cappuccio, L. D'Elia, P. Strazzullo y M. A. Miller, «Sleep Duration and All-Cause Mortality: A Systematic Review and Meta-Analysis of Prospective Studies,» *Sleep*, vol. 33, n.º 5, págs. 585-592, 2010.
- [11] E. B. Leary, K. T. Watson, S. Ancoli-Israel et al., «Association of Rapid Eye Movement Sleep With Mortality in Middle-aged and Older Adults,» *JAMA Neurology*, vol. 77, n.º 10, págs. 1241, 2020.
- [12] J. R. Winer, B. A. Mander, S. Kumar et al., «Sleep Disturbance Forecasts -Amyloid Accumulation across Subsequent Years,» *Current Biology*, vol. 30, n.º 21, págs. 4291-4298, 2020.
- [13] M. M. Ohayon, «Epidemiological Overview of sleep Disorders in the General Population,» *Sleep Medicine Research*, vol. 2, n.º 1, págs. 1-9, 2011.
- [14] M. M. Ohayon y S. Smirne, «Prevalence and consequences of insomnia disorders in the general population of Italy,» *Sleep Medicine*, vol. 3, n.º 2, págs. 115-120, 2002.
- [15] M. M. Ohayon, «Prevalence and comorbidity of sleep disorders in general population,» *La Revue du Praticien*, vol. 57, n.º 14, págs. 1521-1528, 2007.
- [16] R. G. Norman, M. A. Scott, I. Ayappa, J. A. Walsleben y D. M. Rapoport, «Sleep Continuity Measured By Survival Curve Analysis,» *Sleep*, vol. 29, n.º 12, págs. 1625-1631, 2006.
- [17] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. Hu y M. D. Vos, «Detection of REM sleep behaviour disorder by automated polysomnography analysis,» *Clinical Neurophysiology*, vol. 130, n.º 4, págs. 505-514, 2019.

- [18] J. B. Stephansen, A. N. Olesen, M. Olsen et al., «Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy,» *Nature Communications*, vol. 9, n.º 1, 2018.
- [19] J. A. E. Christensen, O. Carrillo, E. B. Leary et al., «Sleep-stage transitions during polysomnographic recordings as diagnostic features of type 1 narcolepsy,» *Sleep Medicine*, vol. 16, n.º 12, págs. 1558-1566, 2015.
- [20] R. Vallat y M. P. Walker, «An open-source, high-performance tool for automated sleep staging,» *eLife*, vol. 10, 2021.
- [21] R. S. Rosenberg y S. V. Hout, «The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring,» *Journal of Clinical Sleep Medicine*, vol. 09, n.º 01, págs. 81-87, 2013.
- [22] M. Younes, J. Raneri y P. Hanly, «Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability,» *Journal of Clinical Sleep Medicine*, vol. 12, n.º 06, págs. 885-894, 2016.
- [23] L. Höller y H. Riemer, «Comparison of Visual Analysis and Automatic Sleep Stage Scoring (Oxford Medilog 9000 System),» *European Neurology*, vol. 25, n.º 2, págs. 36-45, 1986.
- [24] H. Phan y K. Mikkelsen, «Automatic sleep staging of EEG signals: recent development, challenges, and future directions,» *Physiological Measurement*, vol. 43, n.º 4, 2022.
- [25] E. López-Larraz, C. Escolano, A. Robledo-Menéndez, L. Morlas, A. Alda y J. Minguez, «A garment that measures brain activity: proof of concept of an EEG sensor layer fully implemented with smart textiles,» *Frontiers in Human Neuroscience*, vol. 17, 2023.
- [26] K. B. Mikkelsen, Y. R. Tabar, S. L. Kappel et al., «Accurate whole-night sleep monitoring with dry-contact ear-EEG,» *Scientific Reports*, vol. 9, n.º 1, 2019.
- [27] K. S. Park y S. H. Choi, «Smart technologies toward sleep monitoring at home,» *Biomedical Engineering Letters*, vol. 9, n.º 1, págs. 73-85, 2019.
- [28] H.-V. V. Ngo, T. Martinetz, J. Born y M. Mölle, «Auditory Closed-Loop Stimulation of the Sleep Slow Oscillation Enhances Memory,» *Neuron*, vol. 78, n.º 3, págs. 545-553, 2013.
- [29] L. Marshall, H. Helgadóttir, M. Mölle y J. Born, «Boosting slow oscillations during sleep potentiates memory,» *Nature*, vol. 444, n.º 7119, págs. 610-613, 2006.
- [30] A. Holzinger, G. Langs, H. Denk, K. Zatloukal y H. Müller, «Causability and explainability of artificial intelligence in medicine,» *WIREs Data Mining and Knowledge Discovery*, vol. 9, n.º 4, 2019.
- [31] C. S. Nayak y A. C. Anilkumar, «EEG Normal Sleep,» en *StatPearls*. StatPearls Publishing, 2023.
- [32] J. A. Hobson, «A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects,» *Electroencephalography and Clinical Neurophysiology*, vol. 26, págs. 644, 1969.
- [33] C. Iber, S. Ancoli-Israel, A. Chesson y S. Quan, «The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications,» *Westchester, IL: American Academy of Sleep Medicine*, 2007.
- [34] L. Fiorillo, A. Puiatti, M. Papandrea et al., «Automated sleep scoring: A review of the latest approaches,» *Sleep Medicine Reviews*, vol. 48, págs. 101-204, 2019.
- [35] M. K. Pavlova y V. Latreille, «Sleep Disorders,» *The American Journal of Medicine*, vol. 132, n.º 3, págs. 292-299, 2019.

- [36] P. LEMOINE, T. NIR, M. LAUDON y N. ZISAPEL, «Prolonged-release melatonin improves sleep quality and morning alertness in insomnia patients aged 55 years and older and has no withdrawal effects,» *Journal of Sleep Research*, vol. 16, n.º 4, págs. 372-380, 2007.
- [37] R. Heinzer, S. Vat, P. Marques-Vidal et al., «Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study,» *The Lancet Respiratory Medicine*, vol. 3, n.º 4, págs. 310-318, 2015.
- [38] E. S. Arnardottir, E. Bjornsdottir, K. A. Olafsdottir, B. Benediktsdottir y T. Gislason, «Obstructive sleep apnoea in the general population: highly prevalent but minimal symptoms,» *European Respiratory Journal*, vol. 47, n.º 1, págs. 194-202, 2015.
- [39] C. Hublin, J. Kaprio, M. Partinen et al., «The prevalence of narcolepsy: An epidemiological study of the Finnish Twin Cohort,» *Annals of Neurology*, vol. 35, n.º 6, págs. 709-716, 1994.
- [40] O. V. Yakovleva, M. G. Poluektov, O. S. Levin y E. A. Lyashenko, «Sleep and wakefulness disorders in neurodegenerative diseases,» *Zhurnal neurologii i psikiatrii im. S.S. Korsakova*, vol. 118, n.º 4, pág. 83, 2018.
- [41] K. E. Innes, T. K. Selfe y P. Agarwal, «Prevalence of restless legs syndrome in North American and Western European populations: A systematic review,» *Sleep Medicine*, vol. 12, n.º 7, págs. 623-634, 2011.
- [42] A. Biasucci, B. Franceschiello y M. M. Murray, «Electroencephalography,» *Current Biology*, vol. 29, n.º 3, R80-R85, 2019.
- [43] H. Berger, «Über das Elektrenkephalogramm des Menschen,» *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 87, n.º 1, págs. 527-570, 1929.
- [44] M. Schupp y C. D. Hanning, «Physiology of sleep,» *BJA CEPD Reviews*, vol. 3, n.º 3, págs. 69-74, 2003.
- [45] L. D. Gennaro, M. Ferrara y M. Bertini, «The spontaneous K-complex during stage 2 sleep: is it the ‘forerunner’ of delta waves?» *Neuroscience Letters*, vol. 291, n.º 1, págs. 41-43, 2000.
- [46] J. M. Stern, M. Caporro, Z. Haneef et al., «Functional imaging of sleep vertex sharp transients,» *Clinical Neurophysiology*, vol. 122, n.º 7, págs. 1382-1386, 2011.
- [47] D. Moser, P. Anderer, G. Gruber et al., «Sleep Classification According to AASM and Rechtschaffen & Kales: Effects on Sleep Scoring Parameters,» *Sleep*, vol. 32, n.º 2, págs. 139-149, 2009.
- [48] M. TAKAHARA, S. KANAYAMA, H. NITTONO y T. HORI, «REM sleep EEG pattern: Examination by a new EEG scoring system for REM sleep period,» *Sleep and Biological Rhythms*, vol. 4, n.º 2, págs. 105-110, 2006.
- [49] Y. M. Suradi, F. S. Algreeshah, N. Rincon-Flores y S. R. Benbadis, «Mystery Case: Sawtooth waves: Another normal variant not to over-read as epileptiform discharges,» *Neurology*, vol. 84, n.º 12, e87-e88, 2015.
- [50] G. Fu, Y. Zhou, P. Gong, P. Wang, W. Shao y D. Zhang, «A Temporal-Spectral Fused and Attention-Based Deep Model for Automatic Sleep Staging,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, págs. 1008-1018, 2023.
- [51] S. K. Satapathy, D. Loganathan, S. S y P. Narayanan, «Automated Sleep Staging Analysis using Sleep EEG signal: A Machine Learning based Model,» en *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, 2021, págs. 87-96.
- [52] M. Vaezi y M. Nasri, «Application of Heuristic Feature Selection in EEG based Sleep Stages Classification,» en *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, IEEE, 2020, págs. 1-6.

- [53] T. F. Zaidi y O. Farooq, «Automatic Classification of Sleep Stages using EEG Subbands based Time-spectral Features,» en *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, IEEE, 2021, págs. 720-725.
- [54] M. Diych, Y. Li y P. Wen, «EEG Sleep Stages Classification Based on Time Domain Features and Structural Graph Similarity,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, n.º 11, págs. 1159-1168, 2016.
- [55] P. Memar y F. Faradji, «A Novel Multi-Class EEG-Based Sleep Stage Classification System,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, n.º 1, págs. 84-95, 2018.
- [56] M. Čić, J. Šoda y M. Bonković, «Automatic classification of infant sleep based on instantaneous frequencies in a single-channel EEG signal,» *Computers in Biology and Medicine*, vol. 43, n.º 12, págs. 2110-2117, 2013.
- [57] A. R. Hassan y A. Subasi, «A decision support system for automated identification of sleep stages from single-channel EEG signals,» *Knowledge-Based Systems*, vol. 128, págs. 115-124, 2017.
- [58] P. Chriskos, D. S. Kaitalidou, G. Karakasis et al., «Automatic Sleep Stage Classification Applying Machine Learning Algorithms on EEG Recordings,» en *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2017.
- [59] L. Siyuan, L. Jingyuan, G. Hangping y R. Minhua, «Sleep staging prediction model based on XGBoost,» en *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 2021, págs. 350-353.
- [60] C. Guo, F. Lu, S. Liu y W. Xu, «Sleep EEG Staging Based on Hilbert-Huang Transform and Sample Entropy,» en *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, 2015, págs. 442-445.
- [61] E. Alickovic y A. Subasi, «Ensemble SVM Method for Automatic Sleep Stage Classification,» *IEEE Transactions on Instrumentation and Measurement*, vol. 67, n.º 6, págs. 1258-1265, 2018.
- [62] C. Timplalexis, K. Diamantaras e I. Chouvarda, «Classification of Sleep Stages for Healthy Subjects and Patients with Minor Sleep Disorders,» en *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, IEEE, 2019, págs. 344-351.
- [63] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews e Y. Guo, «Mixed Neural Network Approach for Temporal Sleep Stage Classification,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, n.º 2, págs. 324-333, 2018.
- [64] R. Wei, X. Zhang, J. Wang y X. Dang, «The research of sleep staging based on single-lead electrocardiogram and deep neural network,» *Biomedical Engineering Letters*, vol. 8, n.º 1, págs. 87-93, 2017.
- [65] O. Tsinalis, P. M. Matthews, Y. Guo y S. Zafeiriou, *Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks*, 2016.
- [66] A. Vilamala, K. H. Madsen y L. K. Hansen, «Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring,» en *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 2017.
- [67] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen y M. D. Vos, «DNN Filter Bank Improves 1-Max Pooling CNN for Single-Channel EEG Automatic Sleep Stage Classification,» en *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018.

- [68] A. Malafeev, D. Laptev, S. Bauer et al., «Automatic Human Sleep Stage Scoring Using Deep Neural Networks,» *Frontiers in Neuroscience*, vol. 12, 2018.
- [69] S. Biswal, J. Kulas, H. Sun et al., *SLEEPNET: Automated Sleep Staging System via Deep Learning*, 2017.
- [70] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen y M. D. Vos, «Automatic Sleep Stage Classification Using Single-Channel EEG: Learning Sequential Features with Attention-Based Recurrent Neural Networks,» en *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018.
- [71] A. Supratak, H. Dong, C. Wu e Y. Guo, «DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, n.º 11, págs. 1998-2008, 2017.
- [72] A. N. Olesen, P. J. Jennum, E. Mignot y H. B. D. Sorensen, «Automatic sleep stage classification with deep residual networks in a mixed-cohort setting,» *Sleep*, vol. 44, n.º 1, 2020.
- [73] H. Phan, O. Y. Chen, P. Koch et al., «Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning,» *IEEE Transactions on Biomedical Engineering*, vol. 68, n.º 6, págs. 1787-1798, 2021.
- [74] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen y M. D. Vos, «SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, n.º 3, págs. 400-410, 2019.
- [75] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins y M. D. Vos, «XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, págs. 1-1, 2021.
- [76] A. Vaswani, N. Shazeer, N. Parmar et al., *Attention Is All You Need*, 2017.
- [77] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018.
- [78] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., «Improving language understanding by generative pre-training,» 2018.
- [79] A. L. Goldberger, L. A. N. Amaral, L. Glass et al., «PhysioBank, PhysioToolkit, and PhysioNet,» *Circulation*, vol. 101, n.º 23, 2000.
- [80] B. Kemp, A. Zwinderman, B. Tuk, H. Kamphuisen y J. Obery, «Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG,» *IEEE Transactions on Biomedical Engineering*, vol. 47, n.º 9, págs. 1185-1194, 2000.
- [81] C. O'Reilly, N. Gosselin, J. Carrier y T. Nielsen, «Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research,» *Journal of Sleep Research*, vol. 23, n.º 6, págs. 628-635, 2014.
- [82] G.-Q. Zhang, L. Cui, R. Mueller et al., «The National Sleep Research Resource: towards a sleep data commons,» *Journal of the American Medical Informatics Association*, vol. 25, n.º 10, págs. 1351-1358, 2018.
- [83] S. F. Quan, B. V. Howard, C. Iber et al., «The Sleep Heart Health Study: Design, Rationale, and Methods,» *Sleep*, vol. 20, n.º 12, págs. 1077-1085, 1997.
- [84] S. Mousavi, F. Afghah y U. R. Acharya, «SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach,» *PLOS ONE*, vol. 14, n.º 5, 2019.
- [85] A. Supratak e Y. Guo, «TinySleepNet: An Efficient Deep Learning Model for Sleep Stage Scoring based on Raw Single-Channel EEG,» en *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, págs. 641-644.

- [86] H. Seo, S. Back, S. Lee, D. Park, T. Kim y K. Lee, «Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG,» *Biomedical Signal Processing and Control*, vol. 61, págs. 102-117, 2020.
- [87] E. Eldele, Z. Chen, C. Liu et al., «An Attention-Based Deep Learning Approach for Sleep Stage Classification With Single-Channel EEG,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, págs. 809-818, 2021.
- [88] M. Fu, Y. Wang, Z. Chen et al., «Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography,» *Frontiers in Physiology*, vol. 12, 2021.
- [89] H. Phan, K. Mikkelsen, O. Y. Chen, P. Koch, A. Mertins y M. D. Vos, «SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification,» *IEEE Transactions on Biomedical Engineering*, vol. 69, n.º 8, págs. 2456-2467, 2022.
- [90] M. Dutt, S. Redhu, M. Goodwin y C. W. Omlin, «SleepXAI: An explainable deep learning approach for multi-class sleep stage identification,» *Applied Intelligence*, 2022.
- [91] G. Brandmayr, M. Hartmann, F. Fürbass et al., «Relational local electroencephalography representations for sleep scoring,» *Neural Networks*, vol. 154, págs. 310-322, 2022.
- [92] J. Pradeepkumar, M. Anandakumar, V. Kugathasan et al., *Towards Interpretable Sleep Stage Classification Using Cross-Modal Transformers*, 2022.
- [93] W. Zhou, H. Zhu, N. Shen et al., «A Lightweight Segmented Attention Network for Sleep Staging by Fusing Local Characteristics and Adjacent Information,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, págs. 238-247, 2023.
- [94] H. Phan, K. Lorenzen, E. Heremans et al., *L-SeqSleepNet: Whole-cycle Long Sequence Modelling for Automatic Sleep Staging*, 2023.
- [95] M. Kim, K. Jung y W. Chung, «Automatic Sleep Stage Classification Method based on Transformer-in-Transformer,» en *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*, IEEE, 2023.
- [96] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer et al., «Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard,» en *J. Sleep Res.*, vol. 18, n.º 1, págs. 74-84, 2009.
- [97] F. Andreotti, H. Phan y M. De Vos, «Visualising convolutional neural network decisions in automatic sleep scoring,» en *CEUR Workshop Proceedings*, 2018, págs. 70-81.
- [98] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie e Y. Zhang, *SalientSleepNet: Multimodal Salient Wave Detection Network for Sleep Staging*, 2021.
- [99] T. Lee, J. Hwang y H. Lee, *TRIER: Template-Guided Neural Networks for Robust and Interpretable Sleep Stage Identification from EEG Recordings*, 2020.
- [100] I. Loshchilov y F. Hutter, *Decoupled Weight Decay Regularization*, 2017.
- [101] S. Ioffe y C. Szegedy, «Batch normalization: Accelerating deep network training by reducing internal covariate shift,» en *International conference on machine learning*, 2015, págs. 448-456.
- [102] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever y R. Salakhutdinov, «Dropout: a simple way to prevent neural networks from overfitting,» *The journal of machine learning research*, vol. 15, n.º 1, págs. 1929-1958, 2014.
- [103] J. Gehring, M. Auli, D. Grangier, D. Yarats e Y. N. Dauphin, *Convolutional Sequence to Sequence Learning*, 2017.
- [104] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh y D. Batra, «Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,» en *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, págs. 618-626.

- [105] A. Guillot, F. Sauvet, E. H. During y V. Thorey, «Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging,» *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, n.º 9, págs. 1955-1965, 2020.
- [106] S. Khalighi, T. Sousa, J. M. Santos y U. Nunes, «ISRUC-Sleep: A comprehensive public dataset for sleep researchers,» *Computer Methods and Programs in Biomedicine*, vol. 124, págs. 180-192, 2016.

# Apéndices

# Apéndice A

## Gestión del proyecto

### A.1. Herramientas software y hardware

El desarrollo e implementación de los modelos de aprendizaje profundo, así como el soporte para la interpretabilidad y explicabilidad, ha sido llevado a cabo en el lenguaje de programación Python (v3.10.9), junto con las librerías Keras (v2.11.0) y TensorFlow (v2.11.0). También, destacan el uso de las bibliotecas NumPy (v1.24.2) y Scikit-Learn (v1.2.2), así como Matplotlib (v3.7.1) para la visualización e interpretación de los resultados obtenidos. El almacenamiento, administración y compartición del código del trabajo se ha efectuado en la plataforma de desarrollo de software GitLab.

Por otro lado, la ejecución de los algoritmos y de las correspondientes pruebas se ha realizado sobre el sistema operativo Linux Ubuntu 22.04.1 LTS en una GPU NVIDIA GeForce RTX 3080 Ti con CUDA v11.8.0, cuDNN v8.0 y drivers de NVIDIA v520.61.05.

### A.2. Planificación

La distribución de las diferentes tareas principales que han constituido el proyecto desarrollado a lo largo del tiempo ha sido reflejada en el diagrama de Gantt de la Figura A.1.

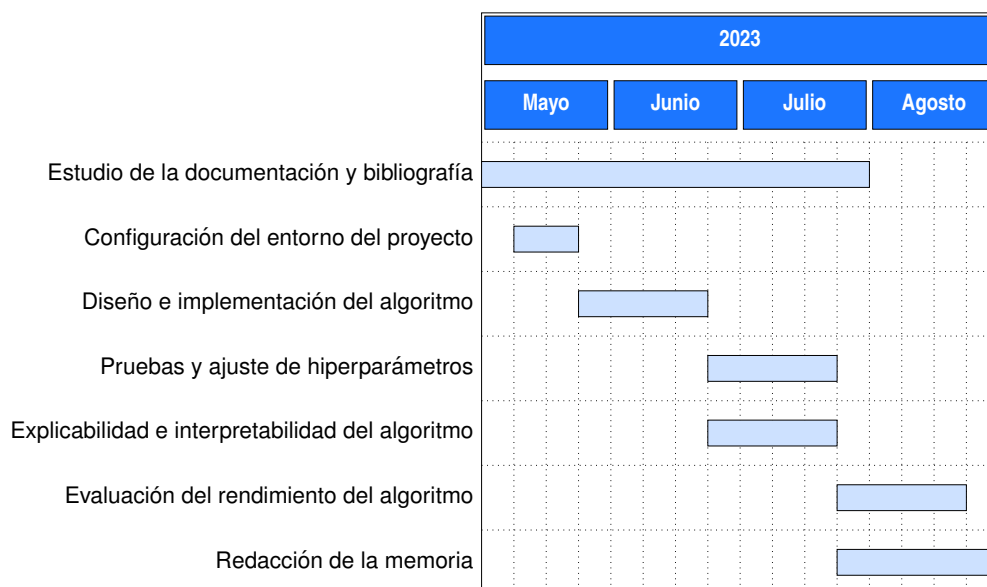


Figura A.1: Diagrama de Gantt del proyecto desarrollado.

## Apéndice B

# Resultados de los métodos de codificación posicional

Con el objetivo de comprobar el rendimiento obtenido con cada técnica de codificación posicional propuesta, se han llevado a cabo 3 pruebas sobre el dataset SleepEDF [79], [80], descrito en el apartado 3.3.1, utilizando la información del canal Fpz-Cz. Para ello, se ha ejecutado el modelo presentado en la sección 3.1 siguiendo la metodología propuesta en el apartado 3.3.2, aplicando un *Leave-One-Out Cross-Validation* con 150 épocas de entrenamiento para cada *fold* y un tamaño de secuencia  $L = 100$ . Cabe destacar que todos los tests se han computado mediante una forma de evaluación *offline*. Por simplicidad, en la Tabla B.1 se muestran los resultados obtenidos respecto a la métrica de *accuracy*.

| Modelo                                      | Accuracy      |
|---------------------------------------------|---------------|
| <i>Sin codificación posicional</i>          | 0.8157        |
| <i>Codificación posicional determinista</i> | <b>0.8413</b> |
| <i>Codificación posicional entrenable</i>   | 0.8247        |

Tabla B.1: Resultados de cada método de codificación posicional sobre el dataset SleepEDF utilizando la información del canal Fpz-Cz.

Por consiguiente, los resultados muestran que la aplicación de cualquier técnica de codificación posicional incrementa el rendimiento del sistema, de forma que el orden de la secuencia de entrada resulta de relevancia en el proceso de clasificación debido a la transición lenta de los procesos fisiológicos que subyacen a las fases del sueño. Sin embargo, el método determinista que utiliza las funciones seno y coseno, correspondientes a las Ecuaciones 3.1 y 3.2, respectivamente, arroja un valor de *accuracy* superior a la técnica entrenable, resultando en la metodología implementada en el modelo de evaluación del sueño.

## Apéndice C

# Estudio de hiperparámetros

A diferencia de los resultados que se obtienen en la mayoría de los datasets ejecutados con los hiperparámetros del modelo (sección 3.1) y la configuración del entrenamiento propuesta por defecto (apartado 3.3.2), se experimenta una caída del rendimiento en el conjunto *Study 3* de Bitbrain, descrito en el apartado 3.3.1, respecto a los resultados esperados en una grabación de estas características, en cuanto a la calidad de la señal y la metodología aplicada en el estudio. Por consiguiente, se ha realizado un estudio de ablación (*ablation*) con el objetivo de incrementar el rendimiento sobre este conjunto de datos, debido a la importancia que presenta al tratarse de grabaciones recopiladas en los laboratorios de Bitbrain a través del dispositivo con el que se pretende integrar el algoritmo. Además, resulta de relevancia estudiar cómo afectan los cambios en los hiperparámetros a la efectividad del modelo. Todas las pruebas planteadas a continuación siguen una forma de evaluación *offline*, y justifican la modificación de algunos aspectos de ejecución de los datasets de Bitbrain respecto a la metodología de evaluación base, apreciables en la Tabla 3.1.

Para ello, en primer lugar, en la Tabla C.1 se puede observar que, en los canales AF7 + AF8 de la diadema textil, el modelo base con la configuración de entrenamiento por defecto arroja una *accuracy* igual al 71.32%, junto con una variabilidad ( $\sigma$ ) elevada. El primer hiperparámetro con el que se experimenta es el tamaño de la secuencia de entrada  $L$  (*batch\_size*), incrementando su valor en 25 épocas hasta 200. Como se puede apreciar, los valores de las métricas computadas aumentan de forma significativa conforme incrementa el valor de  $L$ , determinando el tamaño óptimo de la secuencia de entrada como 200. A partir de este punto, se modifican los valores de otros hiperparámetros de la red (ver Figura 3.1a) respecto al número de cabezas de atención ( $N_H$ ), el número de bloques Transformer ( $N_L$ ) o el número de filtros de las capas convolucionales ( $N_F$ ). Sin embargo, los resultados no experimentan ninguna mejora, determinando que la mejor configuración para este subconjunto se corresponde con los hiperparámetros por defecto y un tamaño de secuencia de entrada  $L = 200$ .

Por otro lado, este efecto también ocurría, en menor medida, en los datos del PSG en los canales F3 + F4, obteniendo un rendimiento del 75.62% en cuanto a la métrica *accuracy* con el modelo por defecto, como se puede apreciar en la Tabla C.2. En base a los resultados anteriores, el valor del tamaño de la secuencia de entrada  $L$  fue incrementado de forma similar, obteniendo un resultado del 0.7705 de *accuracy* para  $L = 175$ , resultando en la configuración de ejecución óptima para este subconjunto.

Finalmente, aunque los resultados en el dataset *Study 1* de Bitbrain se encontraban dentro de lo esperado, resulta fundamental determinar si la modificación de hiperparámetros desemboca en un incremento del rendimiento, tal y como ocurre con los conjuntos anteriores.

Sin embargo, como se puede observar en la Tabla C.3, un aumento en el valor del tamaño de la secuencia de entrada  $L$  no afecta prácticamente a las métricas computadas, manteniendo como óptima la configuración inicial planteada.

| $L$ | $N_H$ | $N_L$ | $N_B$ | $N_F$ | Accuracy                           | Precision     | Recall        | F1-score      |
|-----|-------|-------|-------|-------|------------------------------------|---------------|---------------|---------------|
| 100 | 8     | 4     | 4     | 64    | 0.7132 ( $\sigma = 0.087$ )        | 0.687         | 0.7132        | 0.6804        |
| 150 |       |       |       |       | 0.7512 ( $\sigma = 0.028$ )        | 0.767         | 0.7512        | 0.7471        |
| 175 |       |       |       |       | 0.7753 ( $\sigma = 0.029$ )        | 0.7819        | 0.7753        | 0.7707        |
| 200 |       |       |       |       | <b>0.7785</b> ( $\sigma = 0.026$ ) | <b>0.7865</b> | <b>0.7785</b> | <b>0.7718</b> |
| 175 | 16    |       |       |       | 0.7598 ( $\sigma = 0.051$ )        | 0.7782        | 0.7598        | 0.7582        |
| 200 | 16    |       |       |       | 0.7774 ( $\sigma = 0.027$ )        | 0.7787        | 0.7774        | 0.7708        |
| 200 |       | 8     |       |       | 0.6736 ( $\sigma = 0.026$ )        | 0.5563        | 0.6736        | 0.5939        |
| 200 |       |       | 32    |       | 0.7605 ( $\sigma = 0.108$ )        | 0.7659        | 0.7605        | 0.7493        |

Tabla C.1: Estudio de hiperparámetros sobre el dataset *Study 3* de Bitbrain en los canales AF7 + AF8 de la banda textil. La primera fila se corresponde con el modelo base y las celdas vacías con los parámetros por defecto.

| $L$ | $N_H$ | $N_L$ | $N_B$ | $N_F$ | Accuracy                           | Precision     | Recall        | F1-score      |
|-----|-------|-------|-------|-------|------------------------------------|---------------|---------------|---------------|
| 100 | 8     | 4     | 4     | 64    | 0.7562 ( $\sigma = 0.029$ )        | 0.764         | 0.756         | 0.7509        |
| 75  |       |       |       |       | 0.6243 ( $\sigma = 0.060$ )        | 0.5397        | 0.6243        | 0.5402        |
| 125 |       |       |       |       | 0.7663 ( $\sigma = 0.044$ )        | 0.7835        | 0.7663        | 0.7639        |
| 150 |       |       |       |       | 0.769 ( $\sigma = 0.025$ )         | 0.7746        | 0.7690        | 0.7639        |
| 175 |       |       |       |       | <b>0.7705</b> ( $\sigma = 0.038$ ) | <b>0.7834</b> | <b>0.7705</b> | <b>0.7695</b> |
| 200 |       |       |       |       | 0.7676 ( $\sigma = 0.015$ )        | 0.7689        | 0.7676        | 0.7581        |

Tabla C.2: Estudio de hiperparámetros sobre el dataset *Study 3* de Bitbrain en los canales F3 + F4 del PSG. La primera fila se corresponde con el modelo base y las celdas vacías con los parámetros por defecto.

| $L$ | $N_H$ | $N_L$ | $N_B$ | $N_F$ | Accuracy                           | Precision     | Recall        | F1-score      |
|-----|-------|-------|-------|-------|------------------------------------|---------------|---------------|---------------|
| 100 | 8     | 4     | 4     | 64    | 0.8029 ( $\sigma = 0.043$ )        | 0.8230        | 0.8029        | 0.8029        |
| 125 |       |       |       |       | <b>0.8046</b> ( $\sigma = 0.048$ ) | <b>0.8233</b> | <b>0.8046</b> | <b>0.8040</b> |
| 150 |       |       |       |       | 0.7860 ( $\sigma = 0.063$ )        | 0.8117        | 0.7860        | 0.7865        |
| 175 |       |       |       |       | 0.7958 ( $\sigma = 0.038$ )        | 0.8095        | 0.7958        | 0.7930        |
| 200 |       |       |       |       | 0.7976 ( $\sigma = 0.041$ )        | 0.8066        | 0.7976        | 0.7907        |

Tabla C.3: Estudio de hiperparámetros sobre el dataset *Study 1* de Bitbrain en los canales F3 + F4 del PSG. La primera fila se corresponde con el modelo base y las celdas vacías con los parámetros por defecto.

## Apéndice D

# Resultados con el modelo SleepXAI

Adicionalmente, se ha experimentado con el modelo SleepXAI [90] con el objetivo de evaluar el rendimiento que obtiene sobre algunos de los datasets utilizados en el presente trabajo. La arquitectura de la red, ilustrada en la Figura D.1, se compone de una red CNN para la parte de codificación a nivel de época (Encoder), encapsulada por una capa TimeDistributed de la API de Keras, y una red CNN + CRF (*Conditional Random Field*) para el componente de codificación secuencial (Labeler), a diferencia del modelo propuesto en el apartado 3.1, el cual presenta una arquitectura CNN + Transformer.

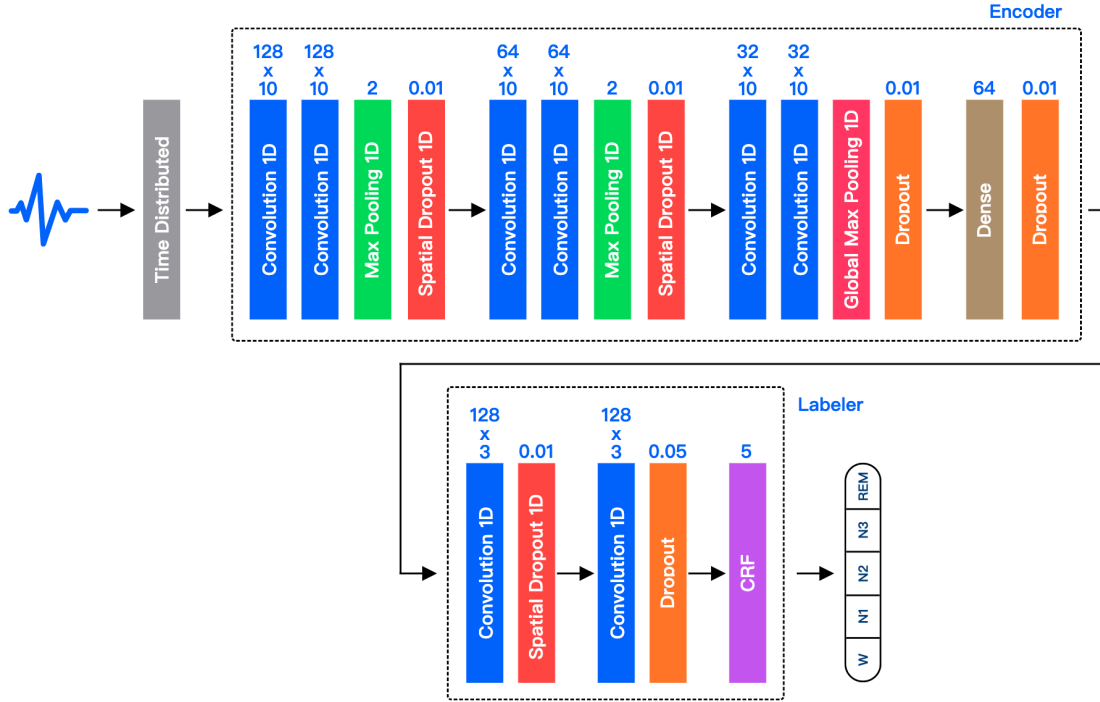


Figura D.1: Arquitectura del modelo SleepXAI.

En cuanto a las pruebas computadas utilizando este modelo, se han planteado varios tests con los datos de Bitbrain y el dataset SleepEDF [79], [80], ambos conjuntos descritos en el apartado 3.3.1. Para ello, se sigue la misma metodología expuesta en la sección 3.1, llevando a cabo un *10-Fold Cross-Validation* con un tamaño de secuencia de entrada  $L = 100$  y 50 épocas de entrenamiento, debido a que la ejecución de este sistema no pudo completarse mediante GPU y fue necesario reducir el tiempo de entrenamiento.

En primer lugar, sobre los conjuntos *Study 1*, *Study 1 consensus* (mismo dataset cuya puntuación es el resultado del consenso de 3 expertos) y *Study 3* de Bitbrain, se expone una evaluación *offline* sobre los datos del PSG en los canales F3 + F4 para analizar el rendimiento que se obtiene con este modelo y compararlo con los resultados de otros algoritmos, así como determinar si la calidad de los datos recopilados es óptima para realizar la clasificación automática del sueño. En esta línea de investigación, los objetivos del experimento se fundamentan en evaluar la efectividad de una red cuya arquitectura difiere de las planteadas anteriormente, comprobar si el rendimiento del *Study 1* es superior al obtenido con el *Study 3*, y si el consenso de varios expertos sobre los mismos datos implica un aumento de la precisión.

Por otro lado, se han planteado 2 pruebas sobre el dataset SleepEDF en los canales Fpz-Cz y Pz-Oz, respectivamente, a través de una evaluación *offline*, de manera que se ejecute la red con el mismo conjunto de datos utilizado en el artículo y se compruebe la similitud en cuanto a los valores de las métricas reportadas por los autores. En la Tabla D.1 se ilustran las pruebas descritas, junto con los resultados obtenidos.

| Test                                                     | Accuracy                    | Precision | Recall | F1-score |
|----------------------------------------------------------|-----------------------------|-----------|--------|----------|
| <i>SXAI00 · Bitbrain · Study 1 · PSG F3+F4</i>           | 0.7754 ( $\sigma = 0.021$ ) | 0.7926    | 0.7754 | 0.7745   |
| <i>SXAI01 · Bitbrain · Study 1 consensus · PSG F3+F4</i> | 0.8373 ( $\sigma = 0.035$ ) | 0.8383    | 0.8373 | 0.832    |
| <i>SXAI02 · Bitbrain · Study 3 · PSG F3+F4</i>           | 0.7275 ( $\sigma = 0.067$ ) | 0.7505    | 0.7275 | 0.7101   |
| <i>SXAI03 · SleepEDF · EEG Fpz-Cz</i>                    | 0.8113 ( $\sigma = 0.049$ ) | 0.8186    | 0.8113 | 0.8064   |
| <i>SXAI04 · SleepEDF · EEG Pz-Oz</i>                     | 0.7849 ( $\sigma = 0.039$ ) | 0.7798    | 0.7849 | 0.7698   |

Tabla D.1: Resultados de la ejecución de las pruebas sobre el modelo SleepXAI.

En cuanto a los resultados obtenidos, se ha determinado, en primera instancia, que el rendimiento arrojado por el modelo SleepXAI es inferior al obtenido con la red propuesta en este trabajo. Por otro lado, el conjunto *Study 1* de Bitbrain experimenta una efectividad superior respecto al *Study 3*, y el dataset *Study 1 consensus* muestra un incremento significativo en los valores de las métricas computadas en referencia al *Study 1*. En consecuencia, la calidad de los datos del conjunto *Study 1* resulta superior a la del *Study 3*, mientras que el consenso de varios expertos en lugar de únicamente uno permite que la clasificación de las mismas grabaciones sea más precisa.

Los resultados del modelo reportados en el artículo original sobre el dataset SleepEDF se corresponden con un valor de *accuracy* igual al 86.8% en el canal Fpz-Cz y 82.8% en el canal Pz-Oz, de manera que difieren en un 5% y un 4%, respectivamente, de los obtenidos a través de las pruebas planteadas, probablemente debido a que la forma de evaluación no es exactamente la misma y por el reducido número de etapas de entrenamiento con las que ha contado el algoritmo para su ejecución. Además, cabe mencionar que, en prácticamente todas las pruebas, las métricas de *accuracy*, *precision*, *recall* y *F1-score* son bastante similares entre sí, y la variabilidad ( $\sigma$ ) respecto a la *accuracy* se encuentra dentro del rango esperado, a excepción del *Study 3*, cuyo valor es del 6.7%.

Finalmente, en la Figura D.2 se encuentran las matrices de confusión para cada uno de los tests realizados, en las que se puede apreciar que las confusiones se concentran en la fase N1 del sueño, la cual es clasificada incorrectamente como la etapa de vigilia (W) en muchas ocasiones, a diferencia del efecto que se supone conseguir en el artículo original aplicando la

capa CRF, la cual pretende mejorar la precisión obtenida en las etapas con menor número de instancias. Por otro lado, las fases W y N2 obtienen unos resultados óptimos y sólidos para todas las pruebas, mientras que las etapas N3 y REM experimentan una variación mayor, siendo ampliamente clasificadas de forma errónea como la fase N2 del sueño.

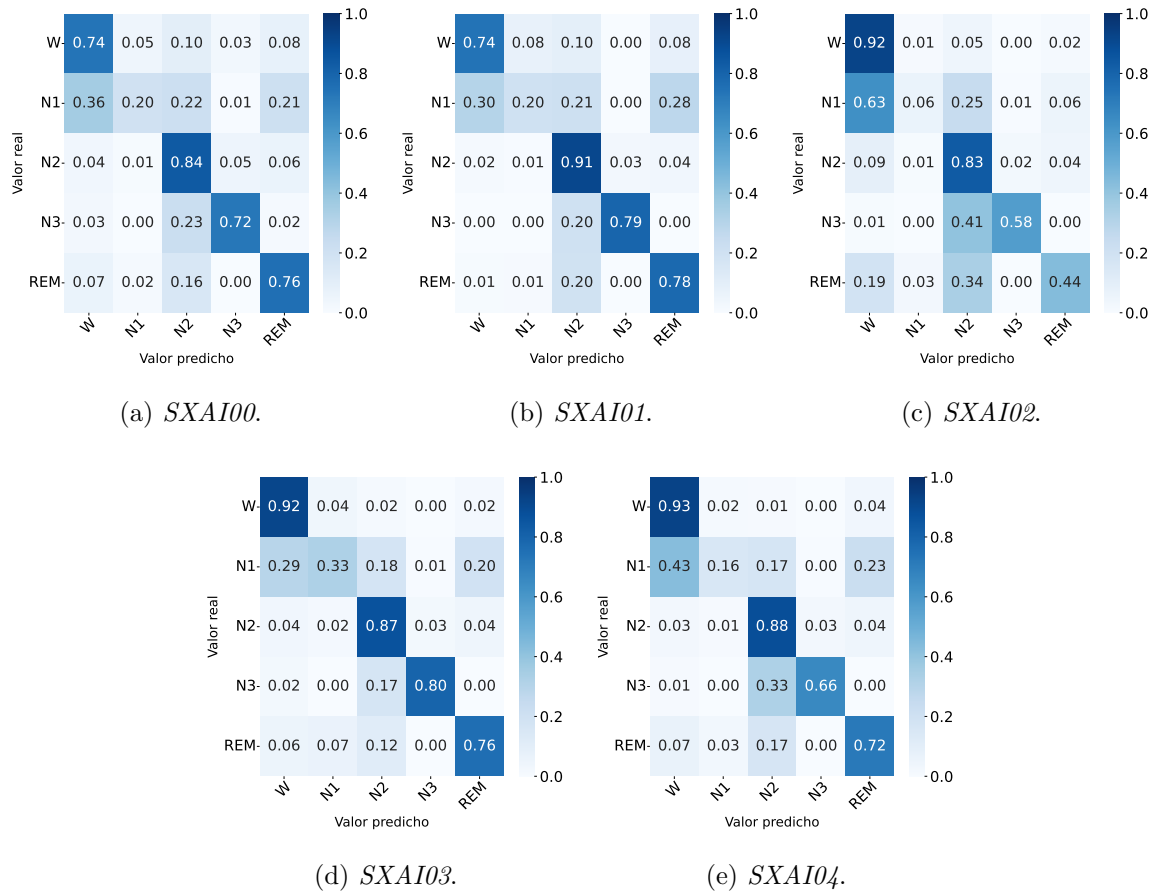


Figura D.2: Matrices de confusión para cada test realizado con el modelo SleepXAI.

## Apéndice E

# Resultados sobre el dataset ISRUC-Sleep

El informe de clasificación obtenido a partir de la ejecución del modelo desarrollado (sección 3.1) sobre el dataset ISRUC-Sleep (descrito en el apartado 3.3.1) siguiendo la metodología de evaluación propuesta (apartado 3.3.2) para las pruebas planteadas (ver Tabla 3.2), se puede apreciar en la Tabla E.1.

En primer lugar, el test *ISRUC00* obtiene un valor de *accuracy* sobre el canal F3-A2 del 66.4% a través de una forma de evaluación *offline*, arrojando un rendimiento superior (+1%) a la evaluación *online* sobre el mismo canal (*ISRUC01*). Posteriormente, las pruebas en los canales F3-A2, C3-A2 y O1-A2, respectivamente, indican que la eficiencia sobre las regiones frontal (65.66%) y central (65.73%) resulta superior a la de la zona occipital (59.85%).

A continuación, el test *ISRUC04* exhibe que la combinación de los canales F3-A2, C3-A2 y O1-A2 provoca un aumento del rendimiento obtenido, superando incluso a la evaluación *offline* sobre el canal F3-A2 (*ISRUC00*) por un 2% de *accuracy*. Adicionalmente, añadir información sobre el EMG, EKG y EOG (*ISRUC05*) no tiene un impacto significativo en los resultados arrojados, manteniendo la efectividad del test anterior.

En general, los valores calculados en todos los tests respecto a las métricas *accuracy*, *precision*, *recall* y *F1-score* son bastante similares entre sí, aunque la precisión tiende a experimentar un resultado ligeramente superior al resto. Estos valores se encuentran por debajo del rango de acuerdo entre evaluadores humanos (70%-80% [21], [22]), constituyendo el dataset sobre el que se ha obtenido una efectividad menor.

| ISRUC-Sleep    | Accuracy                    | Precision | Recall | F1-score |
|----------------|-----------------------------|-----------|--------|----------|
| <i>ISRUC00</i> | 0.6640 ( $\sigma = 0.061$ ) | 0.7287    | 0.6639 | 0.6701   |
| <i>ISRUC01</i> | 0.6566 ( $\sigma = 0.059$ ) | 0.6982    | 0.6566 | 0.6494   |
| <i>ISRUC02</i> | 0.6573 ( $\sigma = 0.064$ ) | 0.6999    | 0.6573 | 0.6455   |
| <i>ISRUC03</i> | 0.5985 ( $\sigma = 0.051$ ) | 0.6348    | 0.5984 | 0.5805   |
| <i>ISRUC04</i> | 0.6827 ( $\sigma = 0.063$ ) | 0.7205    | 0.6826 | 0.6741   |
| <i>ISRUC05</i> | 0.6887 ( $\sigma = 0.094$ ) | 0.7257    | 0.6887 | 0.6802   |

Tabla E.1: Resultados del informe de clasificación sobre el dataset ISRUC-Sleep.

Finalmente, en lo que respecta a las matrices de confusión de la Figura E.1, se puede apreciar que el nivel de confusiones es elevado para prácticamente todas las fases del sueño, exceptuando la etapa de vigilia (W). Al igual que ocurre en el resto de datasets, la fase N1 es la que peores valores arroja, siendo clasificada incorrectamente como las etapas W o N2 en muchas ocasiones, seguida de la fase N2, mientras que las etapas N3 y REM son más similares en cuanto a la precisión obtenida en la mayoría de los casos analizados.

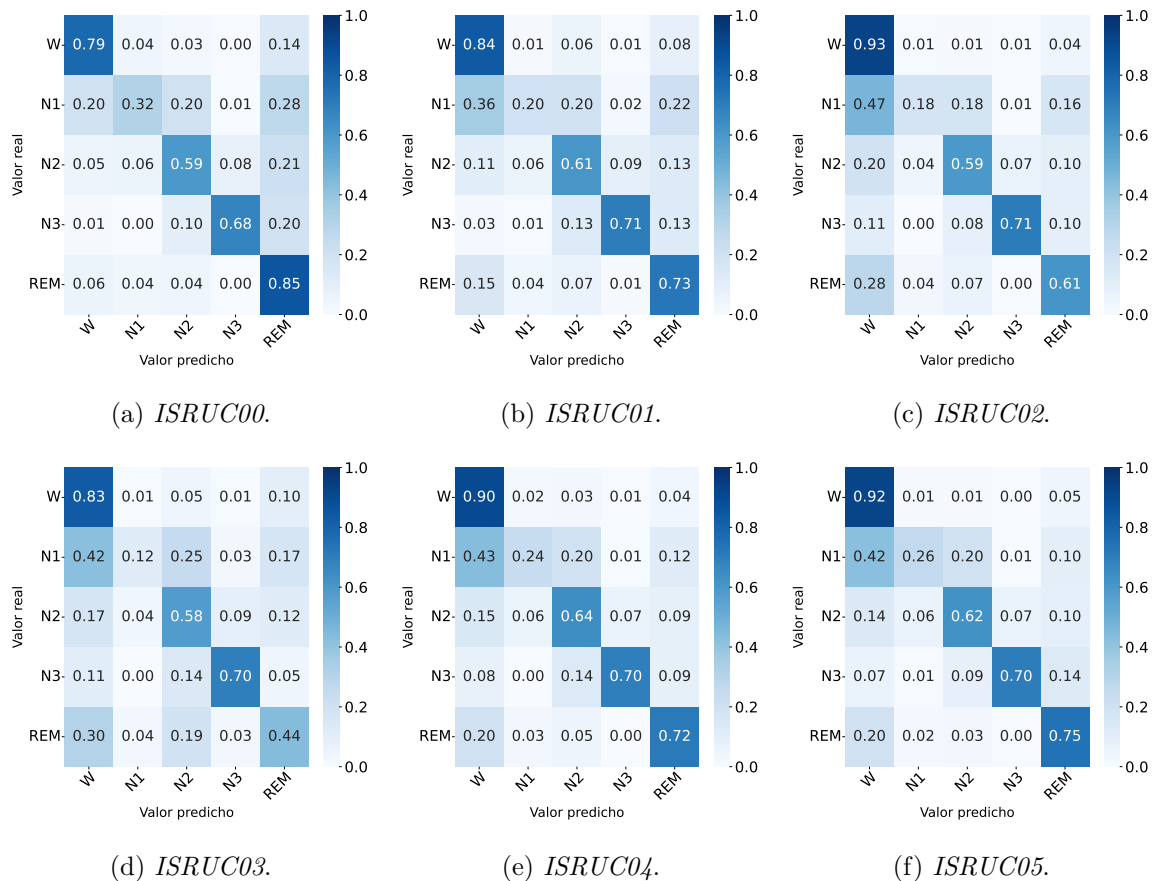


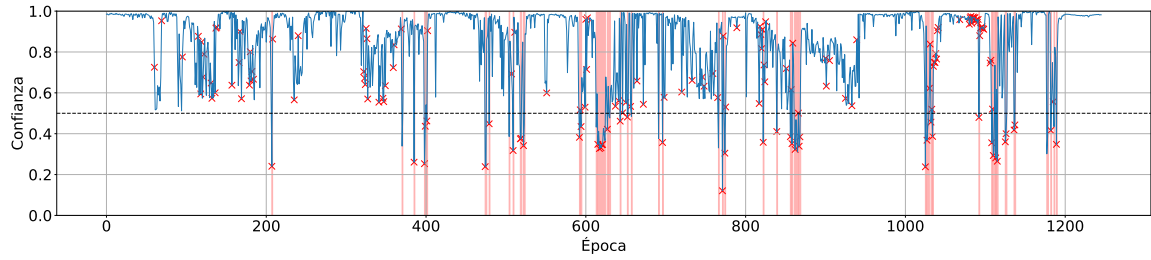
Figura E.1: Matrices de confusión para cada test realizado sobre el dataset ISRUC-Sleep.

## Apéndice F

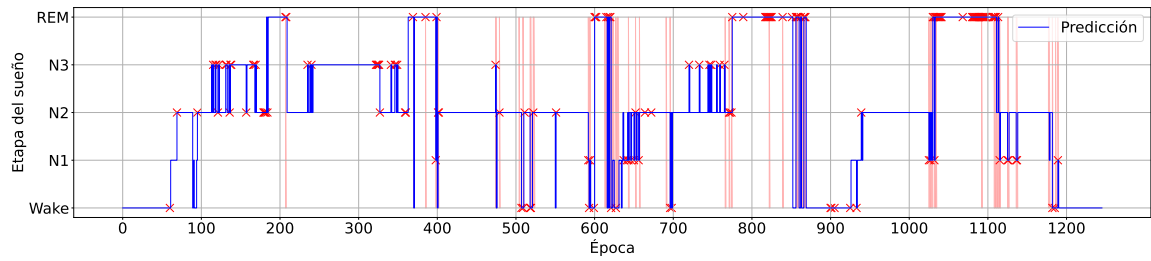
# Resultados adicionales sobre explicabilidad e interpretabilidad

Adicionalmente, se presentan los resultados en cuanto a explicabilidad e interpretabilidad de la ejecución del modelo en el sujeto 5 (SC405) del dataset SleepEDF, utilizando el canal Pz-Oz del EEG y siguiendo una metodología de evaluación *offline*. En consecuencia, en primer lugar, en la Figura F.1 se ilustra el estudio de la confianza del modelo y los hipnogramas generados, en los que se puede observar que los fallos se concentran en zonas con un nivel de confianza bajo y en las regiones de transición entre las diferentes fases del sueño. Además, muchas de las épocas que experimentan una incertidumbre por encima del umbral establecido son clasificadas incorrectamente (concretamente el 64%).

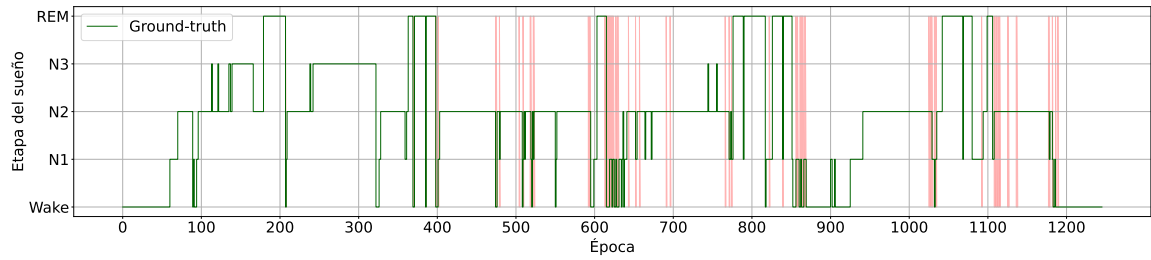
Por otra parte, en la Figura F.2 se muestra el análisis de las probabilidades de salida del modelo, donde se puede apreciar la distribución de los valores de probabilidad de cada época de la grabación sobre el mapa de calor y el gráfico de barras generados. En siguiente lugar, la Figura F.3 ilustra los resultados obtenidos de la aplicación del método Grad-CAM, resaltando las regiones de las épocas que más activación experimentan en relación con las reglas definidas en el manual de la AASM. Finalmente, en la Figura F.4 se puede observar la influencia de las épocas vecinas en la clasificación de la época objetivo respecto a los valores de los pesos de autoatención para algunas épocas de la segunda secuencia ( $L = 100$  épocas) de la grabación correspondiente.



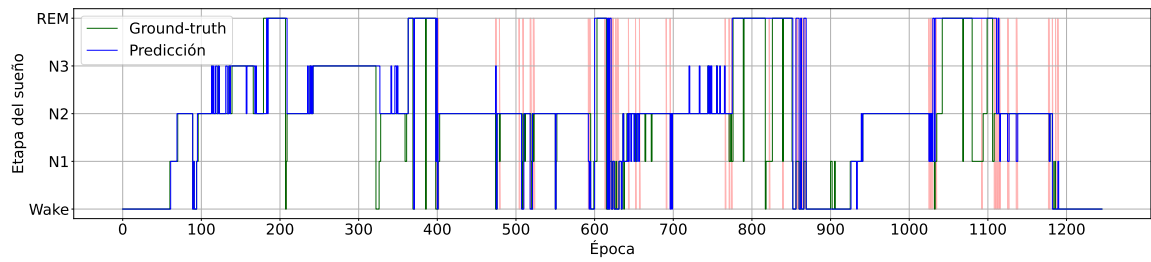
(a) Cuantificación de la confianza, donde  $\times$  representa los fallos ( $threshold = 0.5$ ).



(b) Hipnograma asociado a las predicciones del modelo, donde  $\times$  representa los fallos.

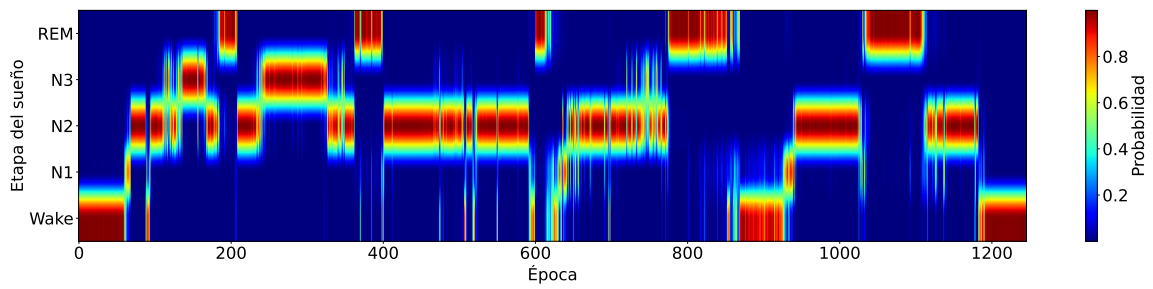


(c) Hipnograma asociado al Ground-truth.

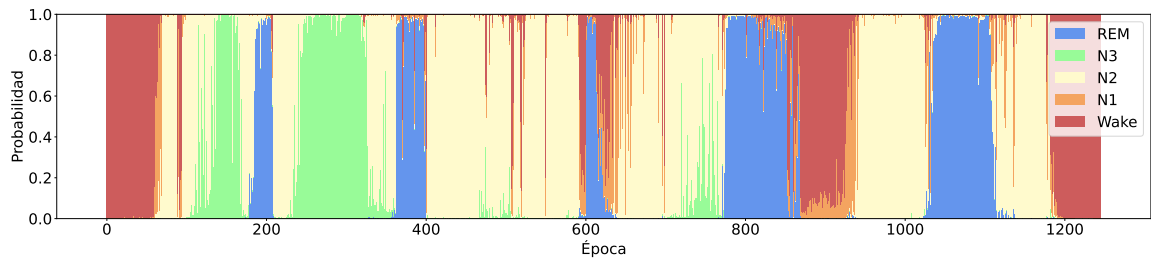


(d) Hipnograma correspondiente a las predicciones del modelo y el Ground-truth.

Figura F.1: Ilustración de la estimación de la confianza del modelo en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( $accuracy = 87.08\%$ , evaluación *offline*).



(a) Mapa de calor asociado a cada época de la grabación en el que se ilustra la intensidad de los valores de probabilidad de las etapas del sueño.



(b) Gráfico de barras correspondiente a cada época de la grabación en el que se muestra un vector de colores según la intensidad de las probabilidades de las fases del sueño.

Figura F.2: Análisis de las probabilidades de salida del modelo en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( $accuracy = 87.08\%$ , evaluación *offline*).

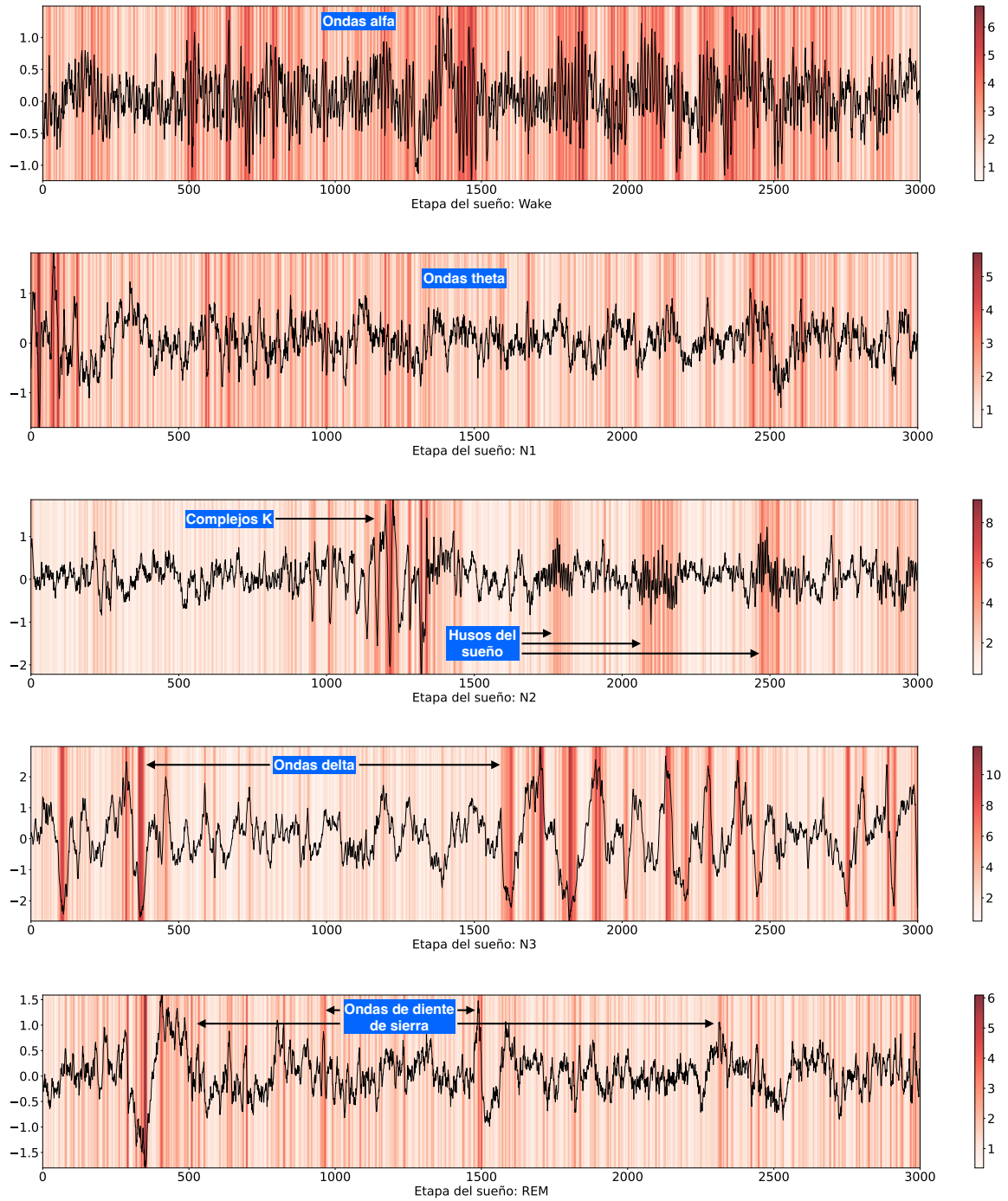


Figura F.3: Resultados obtenidos mediante la aplicación del método Grad-CAM sobre el modelo (segunda capa convolucional de la red), indicando las regiones de las épocas que más activación experimentan en relación con las reglas definidas en el manual de la AASM, correspondientes a cada fase del sueño. La ejecución se ha realizado en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( $accuracy = 87.08\%$ , evaluación *offline*).

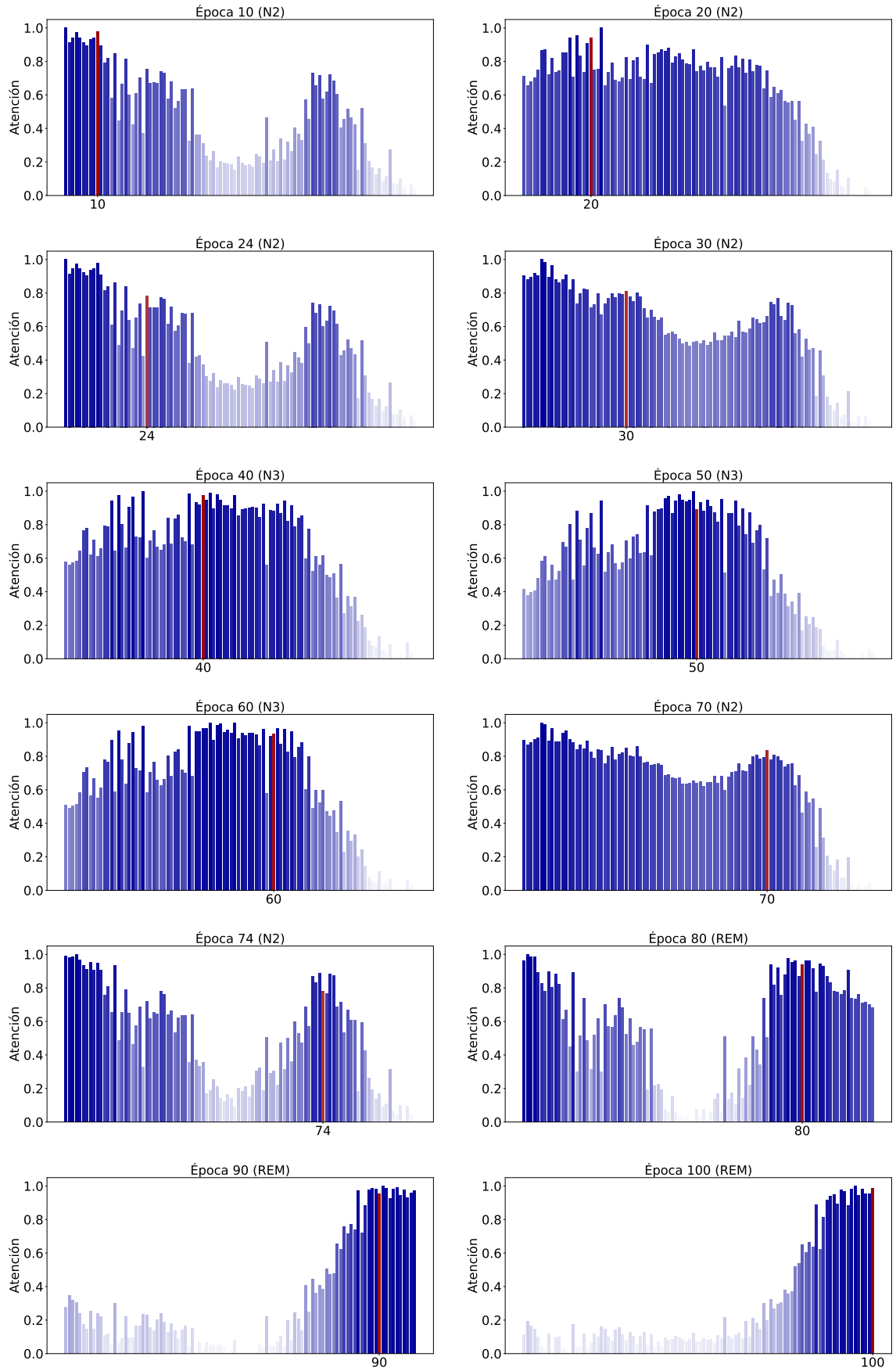


Figura F.4: Resultados del análisis de la influencia de las épocas vecinas respecto a la clasificación de la época objetivo (primera capa Multi-Head Self-Attention de la red) en el sujeto 5 del dataset SleepEDF utilizando el canal Pz-Oz del EEG ( $accuracy = 87.08\%$ , evaluación *offline*).