Universidad Zaragoza
1542

Trabajo Fin de Máster

Título del trabajo:

Desarrollo de herramienta de Aprendizaje Automático para detección de cáncer de pulmón

Autor

Ricardo Teruel Arauzo

Directores

Jacobo Ayensa Jiménez
Manuel Doblaré Castellano

Titulación del autor

Máster en Ingeniería Biomédica

Escuela de Ingeniería y Arquitectura

2019-2021

RESUMEN

El cáncer de pulmón es una de las principales causas de muerte en países desarrollados y su detección precoz ha demostrado ser crítica en la reducción de la mortalidad. Para conseguir detectar los casos de forma temprana, antes de la aparición de síntomas, sería pertinente realizar pruebas a individuos de riesgo, como los fumadores habituales. De las pruebas disponibles, las menos invasivas son las tomas de imagen médica, como la Tomografía Axial Computarizada (TAC). Estas imágenes son revisadas por radiólogos, pero el tiempo del que estos disponen no es ilimitado, especialmente en un contexto de escasez de recursos. Así, herramientas que aceleren este proceso podrían impactar directamente en la detección precoz de esta enfermedad y, por ende, en la reducción de la mortalidad.

El objetivo de este Trabajo de Fin de Máster es desarrollar una herramienta útil en el entorno clínico para detectar el cáncer de pulmón en etapas tempranas de la enfermedad a partir de un TAC de paciente, que sirva de asistencia al diagnóstico, así como para el entrenamiento de futuros radiólogos.

La herramienta consiste en un conjunto de soluciones de Aprendizaje Automático, en este caso, redes neuronales y árboles de decisión. A partir de una imagen TAC de un paciente, la herramienta no sólo da como resultado una clasificación binaria, sino que también aporta una secuencia de características de interés de los nódulos potencialmente cancerígenos para guiar el diagnóstico y ayudar al radiólogo a comprender dicho resultado, huyendo del enfoque de caja negra tradicional asociado a las herramientas de inteligencia artificial.

La herramienta general resultante, aunque no consiguió mejorar respecto a los mejores clasificadores del estado del arte en términos de clasificación binaria global, sí que presenta mejorías en algunas de las partes desarrolladas en el marco del proyecto respecto a la herramienta que sirve como punto de partida de este trabajo. Además, la herramienta se ha diseñado para que el clínico pueda acceder a toda la información intermedia (nódulos candidatos, características morfológicas de los nódulos, malignidad, estadísticas a nivel de paciente) que influya en el diagnóstico, haciéndola completamente transparente y buscando en todo momento la explicabilidad.

Contenido

CAPÍTULO I: Introducción

1.1 – Contexto sanitario

1.1.1 – Cáncer de pulmón: generalidades

El cáncer es un conjunto de enfermedades que se pueden originar en casi cualquier órgano o tejido del cuerpo cuando células anormales crecen descontroladamente, sobrepasan sus límites habituales para invadir partes adyacentes del cuerpo y/o se propagan a otros órganos. Este último proceso se denomina metástasis.

Cuando el cáncer se origina en los pulmones, se denomina cáncer de pulmón. El principal mecanismo etiológico de este tipo de cáncer es la exposición al tabaco, ya sea de manera directa o indirecta. Adicionalmente, puede estar causado por factores ambientales (inhalación de sustancias cancerígenas, contaminación aérea, exposición a radiación), genéticos (predisposición hereditaria) y/o biológicos (infección por virus del VIH). El proceso del cáncer de pulmón es similar al de otros tipos de cáncer. La célula normal que se transforma en la célula tumoral se encuentra en el epitelio que reviste todo el árbol respiratorio desde la tráquea hasta el bronquiolo terminal más fino, y las células que se encuentran en los alveolos pulmonares (Remon, 2019). La proliferación descontrolada de estas células y su acumulación resulta en una masa denominada tumor primario, creando severos problemas a las células y órganos adyacentes. Principalmente, el tumor los priva de recursos y ocupa su espacio, por lo que ven su funcionamiento mermado, como sería el caso de la obstrucción de las vías respiratorias bajas en el cáncer de pulmón. De este tumor primario pueden surgir nuevos tumores metastásicos como consecuencia de la migración de células tumorales a otros órganos o tejidos, típicamente a través del torrente sanguíneo o el sistema linfático. El cáncer de pulmón tiende a diseminarse al cerebro, los huesos, el hígado y las glándulas suprarrenales (Patel, 2021). Los procesos metastásicos más habituales de este tipo de cáncer se muestran en la Figura 1 (Rudin, 2021).

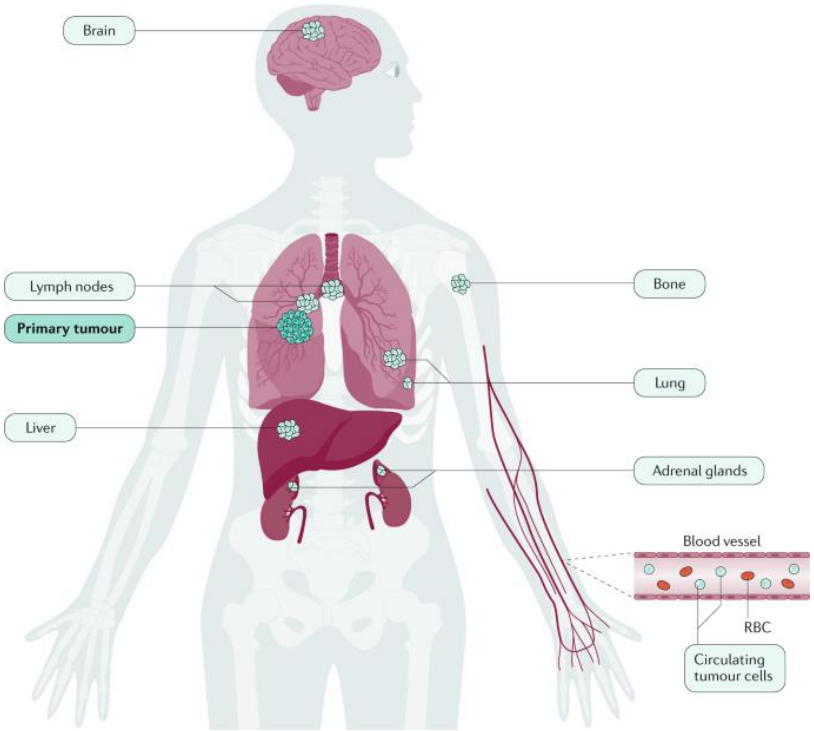


Figura 1. Ilustración del proceso metastásico.

La metástasis es muy prevalente en el cáncer de pulmón y contribuye a un rápido deterioro en la salud del paciente por los mismos mecanismos patofisiológicos que el tumor primario, pero en otras partes del cuerpo (NIH, 2022).

A nivel epidemiológico, el cáncer es la segunda causa de muerte en el mundo; en 2020 causó 9,96 millones de defunciones. Un desglose de esta cifra según el tipo de cáncer puede verse en la Figura 2 (OMS, 2022). En concreto, el cáncer de pulmón afecta actualmente a una de cada dieciséis personas en algún momento de su vida (ACS, 2022) y en 2020, causó 1.8 millones de defunciones.

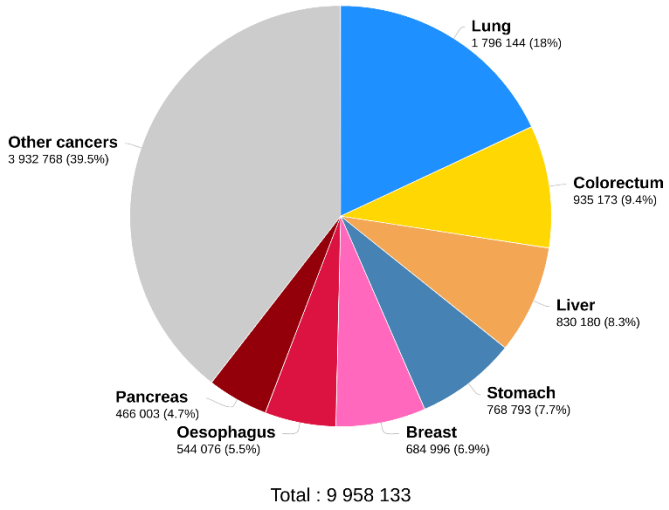


Figura 2. Defunciones desglosadas por tipo de cáncer en el año 2020 (Fuente: (OMS,2022))

En España, se diagnostican cerca de 30.000 casos de cáncer de pulmón cada año, siendo el tercer tipo de cáncer más frecuente en ambos sexos, como ilustra la Figura 3 (SEOM, 2021). En términos de mortalidad, el cáncer de pulmón continúa siendo la primera causa de muerte por cáncer en España también en ambos sexos, con 22.000 víctimas anuales (CEC, 2021).

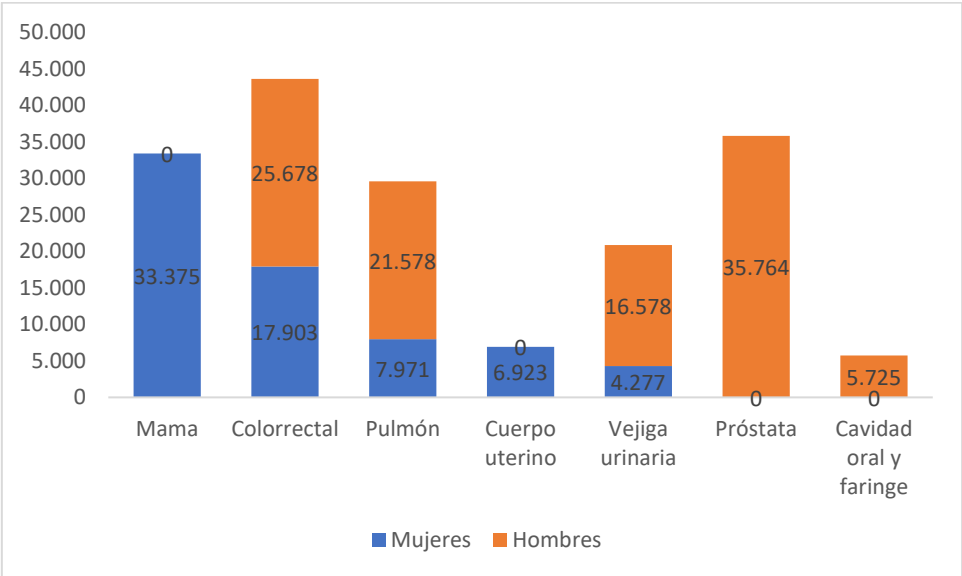


Figura 3. Incidencia estimada de los tumores más frecuentes en hombres y mujeres en España en el año 2021 (SEOM, 2021)

Además del evidente problema de salud individual, el cáncer de pulmón también representa un problema sociosanitario importante en España. Según un estudio de la

Asociación Española Contra el Cáncer (AECC), el coste total estimado de este tipo de cáncer a lo largo de toda la enfermedad por casos diagnosticados cada año asciende a 2.100 millones de euros (Wyman, 2020). Por paciente, el coste asciende a más de 63.000 euros en el caso del cáncer de pulmón local, y a más de 103.000 euros en el caso del metastásico, tal y como se muestra en la Figura 4.

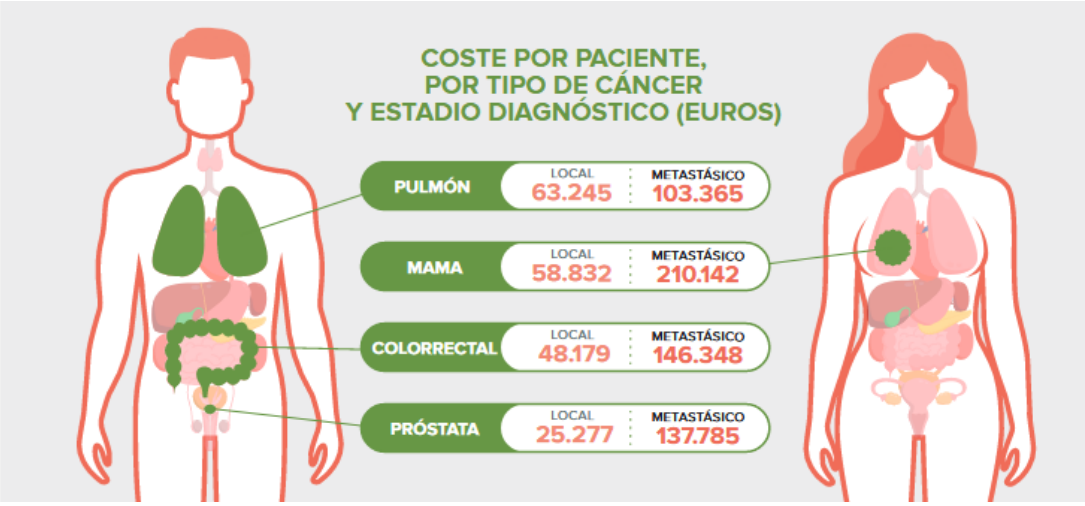


Figura 4.Coste por paciente, por tipo de cáncer y estadio diagnóstico (Fuente: (CEC, 2021))

El tratamiento supone el aspecto más costoso de la enfermedad, seguido de la pérdida de productividad por muerte prematura y de los seguimientos médicos.

1.1.2. – El cribado como estrategia de prevención

Debido al fuerte impacto que tiene el cáncer de pulmón en la sociedad, las posibles soluciones pasan por la prevención y la detección precoz. Así, en cuanto a la primera, se han diseñado distintas estrategias para paliarlo, como las campañas antitabaco o las medidas de control del tabaquismo. La detección precoz se basa en cribados en población de riesgo, es decir, la realización de pruebas diagnósticas sistemáticas a la población sana perteneciente a los grupos poblacionales con más probabilidad de padecer la enfermedad. Una de las pruebas más utilizadas para realizar estos cribados es la tomografía axial computarizada de baja dosis de radiación (en adelante, LDCT, por sus siglas en inglés: *Low-Dose Computed Tomography*). La LDCT suele realizarse en sujetos de entre 55 y 75 años con un consumo de tabaco igual o superior a 30 paquetes al año (CDC, 2021; neumomadrid, 2021).

En este sentido, la detección precoz del cáncer es crucial para la supervivencia, como puede verse en la Tabla 1. La aparición de síntomas, como la tos persistente con esputos de sangre, falta del aliento y dolor de pecho, suele darse cuando el cáncer ya está muy avanzado, así que ha de detectarse mediante otros medios, y es aquí donde entra en juego el cribado.

Etapa de diagnóstico	**Tasa relativa de supervivencia a 5 años**
Localizado	60%
Regional	33%
Distante	6%
Todas combinadas	**23%**

Tabla 1. Tasas relativas de supervivencia a 5 años en función de la etapa del diagnóstico de la enfermedad

En el caso de los Estados Unidos, la *United States Preventive Services Task Force* (USPSTF), recomendó el cribado con LDCT como método para disminuir la mortalidad del cáncer de pulmón basándose en un estudio que demostraba una reducción del 16% en la mortalidad respecto al cribado con radiografía de tórax (Moyer, 2013). Un estudio posterior de los Países Bajos elevó esta disminución al 25% (de Koning, 2020). En cuanto a mejoras absolutas respecto a la ausencia de cribado, si tomamos los datos de España y los comparamos con una tasa de supervivencia del 88% en casos cribados obtenemos un salto del 27% de supervivencia al 88% (Henschke, 2006).

Sin embargo, un problema frecuente de estos cribados es la alta tasa de falsos positivos. Por un lado, en un estudio que recogía los resultados del cribado en 2106 pacientes, 1257 acabaron con resultados positivos. De estos, a 1184 se les recomendaron pruebas sucesivas. En total, solo 73 acabaron siendo sospechosos de cáncer y a 31 de ellos se les acabó diagnosticando en el periodo de un año (Shaughnessy, 2017). Así, un 96% de los individuos con nódulos detectados fueron falsos positivos. Por otro lado, en otro estudio que recoge TACs de 3735 pacientes, solo a 142 de los 905 casos positivos se les acabó diagnosticando cáncer de pulmón, lo que se traduce en una cifra del 84.3% (Hammer, 2020).

Estos falsos positivos pueden acarrear distintos problemas, por ejemplo:

- Carga psicológica en pacientes sanos con consecuencias tan graves como aumento de los suicidios (Lu, 2013; Baade, 2006).
- Complicaciones en pruebas posteriores, como hemorragias o neumotórax en biopsias (Wiener, 2011).
- Utilización de recursos sanitarios en pruebas sucesivas que resultan innecesarias (Álamo-Junquera, 2011).
- Desconfianza de la población en los resultados de las pruebas (Álamo-Junquera, 2011).

Teniendo en cuenta esta información, resulta evidente que la detección temprana del cáncer de pulmón puede salvar multitud de vidas y reducir los costes sanitarios de los tratamientos más agresivos y paliativos. Sin embargo, la detección mediante CT no es un proceso sencillo y requiere de radiólogos expertos. Cuando queremos aplicar este tratamiento a un gran número de pacientes que resulta de un cribado sistemático, la carga de recursos que requerimos del sistema sanitario es muy elevada, y estos podrían dedicarse a otras tareas. Viendo este coste de oportunidad, el cribado masivo ya no resulta tan llamativo. Por tanto, la ayuda a los radiólogos en la toma de decisiones puede resultar fundamental, tanto hoy para la detección, como en el futuro para disminuir el impacto económico y sanitario del cáncer de pulmón.

1.2 – Modelos predictivos en medicina.

En los últimos años, los algoritmos de Aprendizaje Automático (ML por sus siglas en inglés, *Machine Learning*) han impactado enormemente en muchas disciplinas científicas y tecnológicas, entre ellas la medicina preventiva y el diagnóstico médico. Algunos casos de éxito de aplicación de algoritmos de ML para la predicción de enfermedades se han descrito en el cáncer de mama (Houssein, 2021), la diabetes (Kandhasamy, 2015), el Parkinson (Warjurkar, 2021) o la COVID-19 (Ardakani, 2020).

Estos algoritmos pueden alcanzar una alta precisión en el diagnóstico, alcanzando valores de área bajo la curva ROC (AUC por sus siglas en inglés, *Area Under the Curve*), superiores a 0.9. Sin embargo, tienen una importante desventaja: sin una explicación clara de la decisión final que se toma sobre el diagnóstico de la enfermedad, pueden ser muy opacos para el experto (Reddy, 2021).

Este inconveniente es especialmente determinante en el ámbito médico, ya que al estar en juego el bienestar físico y mental de los pacientes, es importante que no exista ninguna clase de incertidumbre en los diagnósticos que dificulte la toma de decisiones de los clínicos sobre el tratamiento de sus pacientes. Por ello, el campo de la ingeniería biomédica está centrando sus esfuerzos en conseguir que los diagnósticos proporcionados por modelos predictivos estén justificados y sean explicables (Collins, 2015).

Otra forma en la que estas herramientas pueden contribuir en la medicina sería la de facilitar el trabajo del profesional sanitario para que sea este quién dé un veredicto, como fue el caso en un estudio donde gracias a una herramienta de *software* basada en ML, el área bajo la curva ROC del desempeño de varios radiólogos aumentó de 0.746 a 0.899 y el tiempo medio de interpretación de las CT se redujo de 168 a 85 segundos (Noguchi, 2022).

1.3 – Objetivo y alcance

El objetivo de este trabajo es crear una herramienta para diagnóstico del cáncer de pulmón a partir de LDCT, que pueda utilizarse en ámbito clínico ya sea como ayuda al diagnóstico o al entrenamiento de nuevos profesionales. Para esto no sólo se busca que clasifique a los pacientes con una precisión aceptable, sino también que pueda proporcionar información de por qué la clasificación ha sido tal y que, en general, sea de utilidad para el médico.

En aras de buscar una mayor interpretabilidad de los criterios de decisión, se busca una herramienta que sea capaz de identificar la posición de los eventuales nódulos malignos, así como de varias de sus características que puedan influir en su diagnóstico (tamaño, posición, forma de las lesiones). Por ello, se trabajará no en la clásica herramienta opaca *end-to-end*, sino en una herramienta que prime la ayuda en la toma de decisiones por parte del radiólogo, aunque arroje un resultado final, en términos de clasificación binaria (o de una probabilidad de salida).

Esta herramienta reduciría en cualquier caso los tiempos de examen de cada paciente y, además de un eventual diagnóstico, podría ayudar a establecer prioridades a la hora de seguir adelante con determinadas pruebas diagnósticas, resultando en un ahorro en tiempo y recursos sanitarios.

Este trabajo toma el relevo de una participación a la *Data Science Bowl 2017*, una competición de ML (kaggle, 2022) donde el objetivo era desarrollar una herramienta de predicción de cáncer de pulmón como la que se busca construir en este trabajo. De este trabajo previo se toman dos de las tres redes neuronales ya entrenadas (en este trabajo denominaremos a dichos modelos DL1 y DL2), que consisten en modelos de detección, y la base de datos que puso a disposición de los concursantes la organización (a partir de ahora nos referiremos a ella como la base de datos DSB), además de algunos fragmentos de código desarrollado.

El primer paso en la realización del proyecto fue la puesta en marcha de la máquina de trabajo (*Workstation*) utilizada, la instalación de entornos de programación y librerías. Una vez en marcha y tras disponer los datos de la base de datos DSB se procedió a ejecutar los códigos, lo cual precisó de su completa comprensión para poder solucionar los distintos errores que surgían. Una vez obtenidos procesados e interpretados los datos otorgados por las dos redes neuronales iniciales, estos se utilizan para entrenar un clasificador, que se diseñará como parte de este trabajo y cuya función será determinar si un paciente padece cáncer o no. Para mejorar este clasificador se aplica una tercera red neuronal, que, entrenada con otra base de datos, asigna un grado de malignidad a un determinado nódulo y se reentrena el clasificador con esta nueva entrada.

1.4 – Estructura del trabajo

Esta memoria se divide en 6 capítulos, cuyo contenido es el siguiente:

- **Capítulo I:** Se introducen conceptos generales que enmarcan la importancia de este trabajo y sus aplicaciones en el ámbito de la medicina. Se discute el impacto de la enfermedad que se quiere combatir, así como la importancia de su correcto diagnóstico y cómo se podría desarrollar una herramienta que incorporase metodologías que ya han probado su eficacia.
- **Capítulo II:** Se describe el estado del arte y las bases de datos que se han utilizado en este trabajo.
- **Capítulo III:** Se aportan descripciones detalladas de cada parte integrante de la herramienta, así como de los elementos más técnicos sobre los aspectos que se trabajan tanto médicos como de ML.
- **Capítulo IV:** Se profundiza más sobre los elementos desarrollados en su totalidad para este trabajo, justificando las decisiones tomadas en su confección.
- **Capítulo V:** Se recogen los resultados de cada parte de la herramienta, así como del desempeño *end-to-end* de la misma, en función de los elementos incluidos.
- **Capítulo VI:** Finalmente se valoran los resultados obtenidos y su utilidad. También se discutirán posibles mejoras o trabajos futuros no contemplados en el marco de este trabajo.

CAPÍTULO II: Estado del arte y bases de datos

En este apartado se presentará el punto de partida, a nivel técnico de este trabajo. Por un lado, el estado del arte, hasta qué punto han llegado otros investigadores trabajando en la solución de problemas similares y, por otro, cuáles son las bases de datos a partir de las cuales se construirá la herramienta.

Para poder discutir los trabajos existentes y compararlos entre ellos se utilizarán unas métricas que se introducirán a continuación. Estas métricas sirven para evaluar cualquier tipo de clasificador binario.

2.1 – Estado del arte

2.1.1. – Algunas métricas habituales para clasificadores binarios

En este capítulo, y en general a lo largo del trabajo, se utilizarán diferentes métricas para evaluar la bondad de una clasificación binaria. A continuación, se describen algunas de ellas:

- **Exactitud:** Es la proporción de resultados correctamente clasificados entre el total de casos examinados. Se denota mediante ACC por sus siglas en inglés (*Accuracy*).

$$\text{ACC} = \frac{VP + VN}{VP + FP + FN + VN}$$

  Donde VP denota el número de verdaderos positivos (pertenecientes a la clase que se busca identificar, como podría ser el número de pacientes con cáncer correctamente clasificados), VN denota el número de verdaderos negativos (no pertenecientes a la clase que se busca identificar correctamente clasificados), FP denota el número de falsos positivos (casos negativos identificados como positivos) y FN el número de falsos negativos (casos positivos identificados como negativos).

- **Razón de Verdaderos positivos**: También conocido como sensibilidad es la porción de VP identificados entre el total de casos positivos se denota mediante TPR por sus siglas en inglés *True Positive Rate*.

$$\text{TPR} = \frac{VP}{VP + FN}$$

- **Razón de Falsos positivos:** También conocido como el complemento de la unidad de la especificidad (1-especificidad) es la proporción de FP entre el total de casos negativos denotado mediante FPR por sus siglas en inglés *False Positive Rate*.

$$\text{FPR} = \frac{FP}{VN + FP}$$

- **Área bajo la curva:** El área bajo la curva ROC (Figura 5) es una métrica que se usa frecuentemente en casos de regresión binaria dado que recoge en un solo índice la sensibilidad a ambas clases. Para obtener la curva ROC (*Receiver Operating Characteristics)* se representan el TPR frente al FPR para cada umbral entre 0 y 1 a partir del cual consideramos un caso como positivo. Los valores del AUC están comprendidos entre 0.5 para un clasificador aleatorio y 1 para un clasificador perfecto.

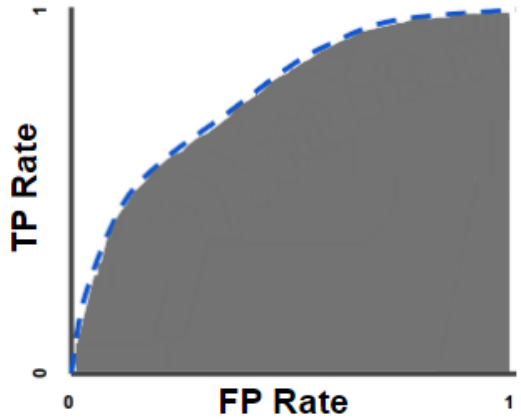


Figura 5. Ejemplo de curva ROC. AUC sombreada

- **LogLoss:** Es otra métrica frecuentemente utilizada para el entrenamiento y evaluación de algoritmos de ML. Esta cuantifica la diferencia entre los valores reales de cada muestra y el valor asignado por el algoritmo, de forma que cuanto más acertadas sean las predicciones menor será el valor de la métrica. Se calcula usando la expresión:

$$\mathrm{LL} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)]$$

Donde N es el número de observaciones, p_i es la probabilidad de pertenencia a la clase positiva asignada a la observación i e y_i su clase real (Dembla, 2022). Los valores de la LL están comprendidos entre $\ln(2) \approx 0.69$ para un clasificador aleatorio y 0 para un clasificador perfecto ya que si $\mathrm{LL} > 0.69$, el clasificador que invierte las clases mejoraría el resultado.

2.1.2 – Revisión bibliográfica

Como se discute el del apartado 1.2, hay un gran número de falsos positivos en las pruebas actuales de cribado de cáncer de pulmón. Si bien solo a la mitad de estos se les requieren pruebas adicionales, pues la otra mitad solo precisa de un seguimiento pasado un periodo de tiempo, las consecuencias ya descritas hacen que esta situación sea un problema a la hora de aplicar cribados de manera sistemática.

En consecuencia, recientemente se han hecho esfuerzos por implementar técnicas de ML en la práctica clínica relacionada con el cáncer de pulmón, entre los cuales este proyecto es uno más. En la literatura existente se describen múltiples herramientas de diagnóstico muy prometedoras, algunas de las cuales se mencionarán a continuación.

En 2017 se llevó a cabo una competición organizada por *Kaggle*, una subsidiaria de Google LLC, llamada *Data Science Bowl 2017* (DSB2017) que ofrecía 1 M$. en premios para aquel equipo que consiguiese discriminar mejor según TACs si un paciente tenía o no cáncer.

En esta competición, los mejores equipos consiguieron AUCs de 0.877 (Intervalo de confianza al 95%: 0.842, 0.910) (equipo grt123), 0.900 (Intervalo de confianza al 95%: 0.870, 0.928) (equipo Aidence) y 0.902 (Intervalo de confianza al 95%: 0.871, 0.932) (equipo JWDH) (Jacobs, 2021). Este trabajo toma como punto de partida algunos elementos de una herramienta desarrollada en dicho concurso (equipo VdP) que obtuvo una LL de 0.46 y una AUC de 0.84.

Posteriormente, otros trabajos han conseguido aumentar estas cifras. Ardila y colaboradores reportan una AUC de 0.96, que desciende hasta el 0.92 si se realiza un seguimiento con imágenes sucesivas. Al tener más información, proveniente de las imágenes de seguimiento, esperaríamos que los algoritmos funcionasen mejor. Sin embargo, hay dos razones para que se dé este caso. La primera, que a los cuadros claros de cáncer no se les toman pruebas de seguimiento, por lo que solo quedan los casos más complicados. La segunda, específica para el modelo, es que el número de pacientes con los que se entrenó el modelo se reducía en este segundo caso (Ardila, 2019).

Llegados a este punto, la pregunta natural es ¿Cuál es el nivel de acierto de un radiólogo? Diversos estudios recogen para los radiólogos, distintos valores de desempeño, pues estos dependerán de la experiencia de estos. Dejando de lado la diferencia de tiempo y el coste de personal de un radiólogo, tenemos AUCs entre 0.92 (Intervalo de confianza al 95%: 0.889, 0.945) (Jacobs, 2021) y 0.846 (Hillis, 2018)

Estas métricas sitúan a los equipos de Aidence y JWDH al nivel de un lector humano con *p*-valores de 0.25 y 0.29 respectivamente, esto quiere decir que las AUCs de estos modelos no eran significativamente peores que aquellas de los radiólogos con los que se compararon mientras que en (Ardila, 2019) se llega hasta a mejorar el rendimiento del radiólogo.

A efectos de comparación, el rendimiento de todos los clasificadores presentados se resume en la Tabla 2:

Clasificador	**AUC**
Equipo VdP	0.84
Equipo grt	0.88
Equipo Aidence	0.90
Equipo JWDH	0.90
Radiólogos	0.92
Ardila	0.96

Tabla 2. Áreas bajo la curva para los diferentes trabajos revisados.

2.2 – Bases de datos

En primer lugar, es conveniente apuntar que los datos, incluso dentro de cada base de datos (BDD), pueden proceder de diversas fuentes, esto hará variar la calidad, tamaño o escala de valores entre las imágenes de unos y otros pacientes. Además, las bases de datos están etiquetadas de distintas maneras, a nivel global de paciente, "Cáncer" / "No Cáncer" o con anotaciones locales como máscaras de pixeles que denoten nódulos y características de estos.

Para este trabajo se han utilizado dos bases de datos:

- La primera es la BDD provista por la DSB, que consiste en TACs hechos a 2058 pacientes de alto riesgo y con nodulación pulmonar, maligna o no, de los cuales se conoce si pasado un año de la exploración fueron diagnosticados con cáncer. Las imágenes vienen en formato DICOM, que es un estándar en imagen médica, que incluye, además de la imagen en sí, una serie de datos que sirven para la identificación del paciente, así como parámetros de configuración de la máquina que tomó las imágenes. Cada imagen contiene una serie de cortes axiales de la cavidad torácica, el número de estos cortes puede variar según la máquina que realiza el escaneo del paciente.

- Además, se utilizó otra BDD, pública, del *Lung Image Database Consortium* e *Image Database Resource Initiative* (LIDC_IDRI). Esta BDD sale de un estudio hecho a 1308 pacientes por siete centros académicos y ocho empresas recogido por el Insituto Nacional del Cáncer en EE. UU. (NCI). De este estudio se tienen los ficheros DICOM de sus TACs de tórax y archivos XML que recogen información aportada por radiólogos como las regiones de píxeles que contienen nódulos, y su malignidad.

CAPÍTULO III: La herramienta, *pipeline*

3.1 – Enfoque y visión de este trabajo

Este trabajo pretende desagregar el flujo completo de trabajo en varias subtareas, tratando de emular el trabajo de un radiólogo. Esto permitirá la creación de una herramienta transparente, con módulos independientes que pueda monitorizarse paso a paso, de forma que se pueda explicar claramente a un clínico y eventualmente a un paciente. Además, en el evento de un fallo diagnóstico, será fácil identificar en que parte hay espacio de mejora.

Para tal efecto, el flujo de trabajo de la herramienta será el siguiente:

- **Preproceso**: Regularización y homogeneización de las imágenes, compresión almacenamiento y segmentación de la parte interna del pulmón.
- **Generación de Regiones de Interés (ROI)**: Identifica las regiones candidatas a ser nódulos.
- **Identificador de nódulos**: Determina en cuales de estas ROI hay, efectivamente, un nódulo, utilizando dos modelos de *Deep Learning* que se detallarán posteriormente.
- **Agregador de nódulos**: determina que conjuntos de cortes están asociados a un nódulo.
- **Extracción de características morfológicas, de textura y malignidad**: se usan herramientas de visión por computador (CV *Computer Vision*) para caracterizar cuantitativamente los nódulos en términos de forma y textura. Además, otro modelo de *Deep Learning* basado en un etiquetado de patólogos les asigna una determinada malignidad.
- **Agregador por paciente**: se pasa de una serie de características a nivel de nódulos a características a nivel de paciente.
- **Clasificador**: en base a las características del paciente se clasifica este en "cáncer"/"no cáncer".

Este flujo de trabajo queda resumido en la Figura 6, en azul se destacan las componentes que se han desarrollado íntegramente en este trabajo.

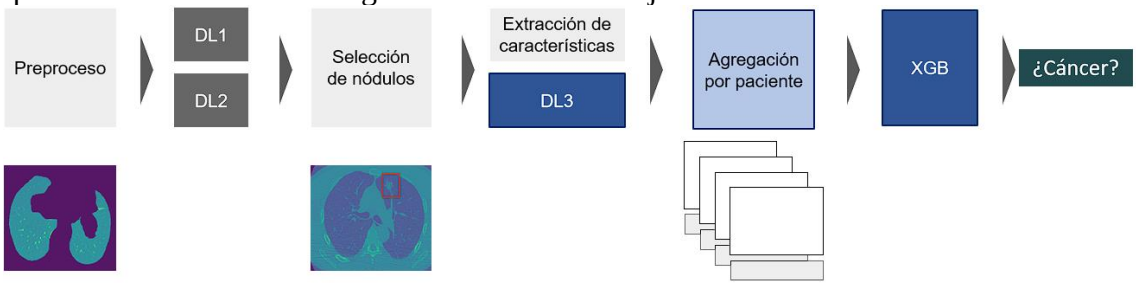


Figura 6. Flujo de trabajo de la herramienta.

3.2 – Descripción detallada de las componentes.

3.2.1 – Preproceso

La generación de imágenes TAC se lleva a cabo con unas máquinas cuyos parámetros de configuración pueden provocar disparidad en los datos, las máquinas pueden tener

distinta resolución o estar calibradas de forma diferente, por ejemplo. Además, los ficheros DICOM son muy pesados debido a como se almacena la información, por lo que es conveniente extraer la información que nos interesa de estos en un formato más manejable. Por último, dentro de cada TAC no solo tenemos información del pulmón, sino también de los tejidos que lo rodean e incluso el aire que queda alrededor del paciente. Es conveniente, para ahorrar posibles conflictos y centrar el poder computacional, discernir entre la región de estudio (el pulmón) y lo demás.

Con el fin de poder utilizar imágenes de bases de datos diversas para entrenar las redes incluidas en la herramienta, así como para clasificar nuevos pacientes, en primer lugar, se requiere una etapa de preprocesado que homogeneice todas las imágenes.

En este preprocesado se cambiará la escala de las imágenes para que el espacio entre píxeles sea igual en todos los pacientes (re-escalado) a base de extraer la información de los archivos DICOM. También se reajustarán las imágenes para que todas ellas posean las mismas dimensiones (512×512 píxeles). También se modifican las imágenes para que todas ellas tengan el mismo espaciado entre cortes axiales.

Para entender los próximos pasos, es conveniente introducir el concepto de Unidades Hounsfield (UH) y dar unas pinceladas sobre como las máquinas generan imágenes TAC de los pacientes.

Una vez el paciente se encuentra recostado dentro de la máquina, una fuente de rayos X móvil gira alrededor de este lanzando haces que serán recibidos por los detectores que se encuentran opuestos a la fuente. Una vez se completa una revolución la máquina avanza en la dirección axial para realizar otro corte. Para cada uno de estos cortes se usan técnicas matemáticas que calculan cual ha sido la absorción en cada punto en el espacio del interior de la cámara, estos valores de absorción se representan en una escala, la escala Hounsfield, donde 0 UH equivalen al agua y -1000 UH al aire a presión y temperatura ambiente (NIBIB, 2022). Cada material tiene un valor de UH característico. En las TAC se puede usar un llamado material de contraste, de aplicación intravenosa que tiene valores especialmente altos, los valores típicos para materiales de interés en este trabajo se recogen en la Tabla 3.

Sustancia u órgano	**Valor típico [UH]**
Aire	-1000
Pulmón normal	(-700,-900)
Grasa	-100
Agua	0
Sangre	50
Musculo	50
Contraste intravenoso	300
Hueso	>1000

Tabla 3. Unidades Hounsfield típicas para distintos materiales presentes en una TAC de pulmón

La distribución típica de valores de píxel de un paciente se puede ver en la Figura 7. El rango de valores de interés estaría comprendido entre los valores 50 y 500, como se puede

apreciar, la cantidad de píxeles que determinarán el estado patológico del paciente son una minoría frente al conjunto de la imagen.

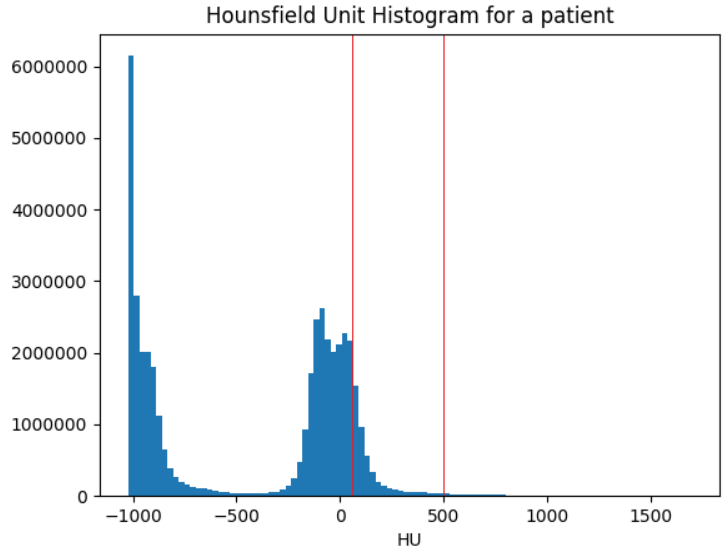


Figura 7. Histograma de valores de píxel [UH] en un paciente

Una vez introducido este concepto continuamos con el funcionamiento de la fase de preprocesado con la segmentación del pulmón. Este proceso consiste en:

- Obtener una imagen con valores de píxel binarios, aplicando simplemente un umbral a -320 UH, lo cual separara el aire que envuelve al paciente y la matriz de los pulmones del resto de tejidos.
- Como sabemos que la pared de los pulmones es relativamente continua y convexa utilizamos un método que "dilata" la región deseada (los pulmones). Esto se hace sencillamente asignando a cada píxel como el mínimo valor de entre los pixeles adyacentes a este. De esta forma tenemos ahora dos regiones. Una de las regiones contiene los pulmones y aire que rodea al paciente, junto con los pixeles más superficiales del cuerpo y el tejido que rodea a los pulmones. La otra el resto de los tejidos del paciente.
- Para separar los pulmones del aire del entorno se toman como referencia las ocho esquinas del TAC, pues estas serán siempre aire, y toda la región continua que las contiene se descarta.
- Finalmente se recorren todos los cortes y para cada corte se seleccionan las dos componentes conexas más grandes asumiendo que dichas regiones se corresponden con los pulmones.

3.2.2 – Generación de regiones de interés

Finalmente, dentro del pulmón se detectan todas las zonas que contengan valores significativamente altos (es la media de valor de píxel en la región abarcada por la segmentación) y se define un recuadro a su alrededor. Los grupos de píxeles delimitados por cada recuadro constituirán las ROI sobre los que actuarán los modelos de detección de nódulos (Figura 8).

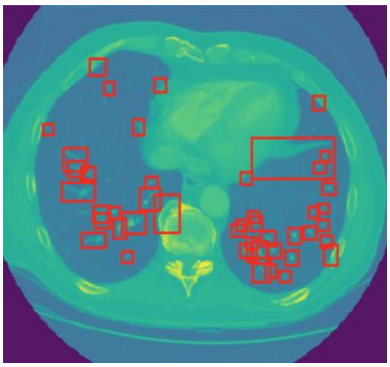


Figura 8. Ejemplo de generación de ROI dentro de un corte.

3.2.3 – Modelo de aprendizaje profundo para detección de nódulos (DL1)

Esta red detectará cuáles de las regiones de interés propuestas se podrían considerar nódulos. Esta red está montada de tal forma que tiene una sensibilidad (*sensitivity*) muy alta, con lo cual es posible que clasifique como nódulos muchas estructuras que realmente no lo sean.

Para esta tarea, se usa una ResNet50, cuyo nombre proviene de *Residual Network* seguido del número de capas que contiene. La elección de este tipo de arquitectura se basa en que las redes neuronales profundas son útiles para encontrar características de distintos niveles de abstracción. Mientras que una red (convolucional) poco profunda solo puede comparar los píxeles más cercanos entre sí, una red neuronal más profunda es capaz de encontrar características que relacionen pixeles que se encuentren a mayor distancia entre sí, así como relaciones adicionales más complejas entre ellos.

Sin embargo, existe una limitación importante a la hora de entrenar estas redes neuronales profundas: cuando el número de capas ocultas es elevado, la complejidad de los cálculos necesarios para reducir los errores en la red (optimización de gradientes) aumenta progresivamente, hasta el punto de llegar a lo que se conoce como "*vanishing or exploding gradients"*, donde pequeñas diferencias en los pesos de la red crean diferencias muy grandes en las conexiones neuronales de las capas iniciales o viceversa (Philipp, 2018).

Para circunvalar este problema, la solución otorgada por las ResNets es crear conexiones entre capas, como se muestra en la Figura 9 que sirvan de atajo para facilitar los cálculos de gradientes.

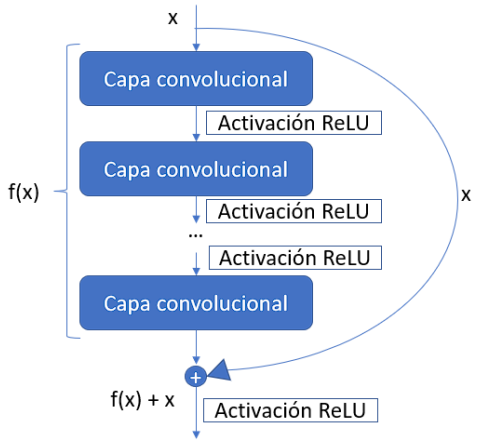


Figura 9. Representación gráfica del funcionamiento de las ResNets

La entrada a esta red neuronal consistía en parches de dimensiones 40×40×3 (40 píxeles de ancho y largo, y 3 de altura) y su salida una clasificación binaria que responde a si el parche examinado contiene o no un nódulo (Figura 10). Esta red se entrenó usando la BDD LIDC-IDRI, ya que esta contiene ejemplos de nódulos etiquetados, a diferencia de DSB que tiene etiquetas a nivel de paciente.

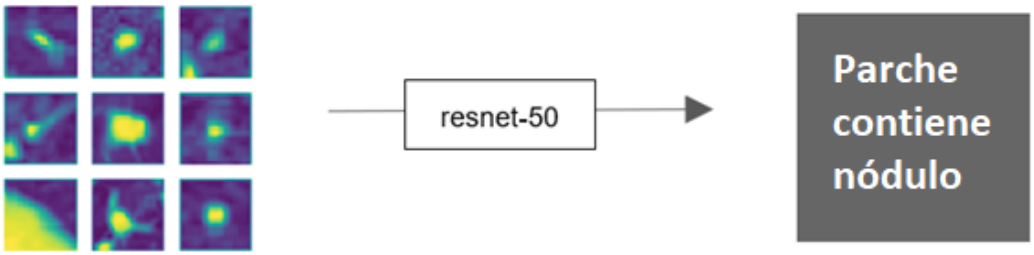


Figura 10. Esquema de funcionamiento del modelo DL1

3.2.4 – Modelo de aprendizaje profundo para la mejorar la especificidad (DL2)

De entre los nódulos propuestos por la red DL1, es necesario discernir cuáles lo son realmente. La red previa está diseñada para tener una sensibilidad muy alta, pero a costa de ello se tiene una especificidad muy baja. Dado esto, un gran número de candidatos serán realmente falsos positivos, de la misma forma que el triaje proporciona falsos positivos. Este problema es muy recurrente en el reconocimiento de objetos de interés en una imagen. Es por ello por lo que hay técnicas ya reconocidas para mejorar el desempeño de este tipo de redes. Una de estas técnicas se conoce como *Hard Negative Mining*.

Gran parte de las ROI marcadas por el proceso descrito en el apartado 3.2.3 corresponden a vasos sanguíneos. Hay ciertas imágenes, por ejemplo, las que tengan vasos que atraviesen el corte perpendicularmente, donde es más difícil discernir si la forma circular resultante es un vaso o un nódulo. Si la red no se entrena específicamente para distinguir dentro de este último tipo de casos, la especificidad será muy baja.

El *Hard Negative Mining* consiste en entrenar una red con ejemplos que previamente ha clasificado erróneamente como positivos para que aprenda a diferenciar los nódulos reales de aquello que se les parezca. El resultado de este entrenamiento es un *score* que determinará la confianza que tiene la red de que lo que haya en la imagen sea realmente un nódulo.

En cuanto, a arquitectura, esta red es igual que la anterior en el sentido de que es una ResNet50 y las entradas son parches 40×40×3, concretamente los que se clasificaron como nódulos con el modelo DL1, ahora en cambio la salida en vez de interpretarse como una clasificación binaria (nódulo / no-nódulo) se conservará la variable de regresión (*score*). Esta red también se entrenó con la BDD LIDC, pero únicamente considerando aquellas regiones que el modelo DL1 clasificaba con bastante certeza como positivos (score mayor que 0.7).

3.2.5 – Agregador de nódulos

A partir de un listado de parches ya filtrados según su puntuación del modelo DL2, vamos a generar un nuevo listado, esta vez de nódulos, tridimensionales, uniendo en cada nódulo los parches en los que aparece, pero tratando de no juntar nódulos cercanos en uno solo. El procedimiento por seguir consistirá en:

- Para cada paciente se carga un parche, los parches se cargan en orden de corte, se comprueba si este puede coincidir con uno de los nódulos ya empezados, si no se crea un nuevo nódulo al que poder añadir futuros parches.
- Para comprobar si coincide con un nódulo existente se calcula la distancia en el plano entre los parches y se divide entre la suma de diámetros, si es mayor a 1 no hay intersección, se establece el límite en 0.5, ya que, de ser mayor, aunque exista intersección no es concéntrico y por tanto se considerarían nódulos cercanos, pero no el mismo. También se calcula la distancia en el eje axial, si es mayor a ocho milímetros se descarta la pertenencia al mismo nódulo.
- Cuando un nódulo lleva abierto y no se le añadido un nuevo parche en 5 cortes se cierra. Al acabar los parches de un paciente se cierran todos los nódulos.
- A cada nódulo se le asigna el mayor diámetro y *score* entre los parches que lo conforman y se le asigna la característica Δz como la distancia máxima entre dos parches de este.

3.2.6 – Extracción de características morfológicas y de textura

En este momento, tenemos uno varios candidatos a nódulo correspondientes por cada paciente, y por cada uno de ellos tenemos una puntuación relativa a cuán probable es que este candidato sea en realidad un nódulo, que no es otra cosa que la salida del modelo DL2. Asumiendo que los nódulos cuya puntuación supera el 0.7 lo son realmente, queda valorar si dichos nódulos son malignos o no, es decir, valorar el estado patológico del paciente.

Consideremos un determinado nódulo. Para saber si este nódulo se corresponde con un nódulo maligno se ha decidido seguir un enfoque que emule el análisis que realizaría un radiólogo. Para ello, se consideran una serie de características morfológicas y de textura que deberían ser propias de nódulos malignos, propuestas por los radiólogos especializados. Dichas características, que pueden cuantificarse usando herramientas de CV se presentan a continuación:

- **Diámetro ($\boldsymbol{D}$)**: diámetro del nódulo en centímetros.
- **Localización vertical del nódulo ($\boldsymbol{z}$)**: número relativo al corte donde aparece el nódulo. Es la altura aproximada a la que se encuentra.
- **Espesor vertical del nódulo ($\boldsymbol{\Delta z}$)**: número de cortes a través de las cuales se extiende el nódulo verticalmente. Esencialmente es la dimensión del nódulo en la dirección axial.
- ***Score* ($\boldsymbol{s}$)**: puntuación otorgada por la DL2, que puede ser interpretada como la probabilidad de que sea efectivamente un nódulo.
- **Posición del nódulo ($\boldsymbol{x, y}$)**: posición del nódulo en los ejes x e y.
- **Autovalores del tensor de inercia ($\boldsymbol{\lambda_1, \lambda_2}$):** Son los valores propios del tensor de inercia del nódulo y codifican la distribución de la masa de un objeto (Gruy, 2015). Entre otros, se verán afectados por la elongación del nódulo, su planitud o su compactación. Esta característica morfológica se puede ver usada en estudios de los mismos nódulos pulmonares (Gu, 2019).
- **Excentricidad ($\boldsymbol{e}$)**: La excentricidad se define como la distancia entre el centro y los focos de una elipse dividida entre la longitud del eje mayor. Aplicado aquí, se aproximaría la máscara del nódulo a una figura elipsoidal. Un nódulo con excentricidad igual a cero sería perfectamente circular y a mayor excentricidad, será más alargado. Hay muchas referencias bibliográficas donde se estipula que los nódulos excéntricos tienden a ser malignos.

  $$e = \sqrt{1 - \frac{{\lambda_1}^2}{{\lambda_2}^2}}$$

  Siendo λ_1 el menor autovalor y λ_2 el mayor.
- **Extensión ($\boldsymbol{A_{\mathrm{ch}}}$)**: Área delimitada por la envolvente convexa del nódulo. Es alta en dos casos: nódulos irregulares con bordes espiculados, los cuales son muy sospechosos de ser malignos; o nódulos muy grandes y regulares, los cuales por esta segunda cualidad no serían sospechosos en sí, pero lo serían dado cierto diámetro. Una extensión alta corresponderá con nódulos malignos.
- **Área ($\boldsymbol{A}$)**: Píxeles cubiertos por el nódulo. Generalmente, un nódulo grande es más sospechoso de ser maligno, pero hay otros factores como los mencionados en el apartado anterior donde un nódulo que captaría la atención del radiólogo no tiene un área especialmente grande.
- **Perímetro ($\boldsymbol{P}$)**: Perímetro que define el nódulo. Esta característica es importante, ya que está relacionada no solo con el tamaño del nódulo (a mayor tamaño, mayor perímetro), sino que también está relacionada con su irregularidad (a mayor irregularidad, mayor perímetro). Ambos parámetros están relacionados con la malignidad del nódulo de la misma forma, de modo que fijándonos en el perímetro detectaremos tanto nódulos grandes como nódulos no tan grandes pero irregulares, y ambos son de interés diagnóstico.
- **Solidez (sl)**: Se define como

  $$sl = \frac{A}{A_{\mathrm{ch}}}$$

  donde A es el área y A_{ch} es la extensión. Cuanto menor este valor la forma es más irregular y por tanto más sospechoso el nódulo de ser maligno.

- **Intensidad media/máxima ($\boldsymbol{I}_{\mathbf{mean}}/\boldsymbol{I}_{\mathbf{max}}$)**: Valor medio y máximo de HU de los píxeles que conforman el nódulo. Al tener las lesiones malignas una vascularización mayor, una CT con contraste como las tratadas en este trabajo, tendrá valores mayores que las lesiones benignas. Un valor de corte para separar lesiones benignas y malignas sería 30HU (Macura, 2012). Por otro lado, la calcificación de los nódulos también aumentará la intensidad con la que estos aparecerán en las imágenes (Khan, 2011), con intensidades de alrededor de 160 HU. Mientras que los valores altos atribuidos a la alta vascularización serían una señal de malignidad, la calcificación suele indicar que la lesión es benigna. La combinación de ambos valores, dividiendo la intensidad media por la máxima, da como resultado la homogeneidad.
- **Homogeneidad (cv)**: Se define como

$$\mathrm{cv} = \frac{I_{\mathrm{mean}}}{I_{\mathrm{max}}}$$

  La calcificación que consideraríamos benigna suele estar presente de forma homogénea por todo el nódulo, a diferencia de la vascularización, que tiene formas más irregulares. De esta forma, la homogeneidad del nódulo debería ayudarnos a clasificar la malignidad de las lesiones, siendo más malignas las que sean menos homogéneas.
- **Posición vertical del nódulo ($\boldsymbol{\hat{z}}$)**: No solo la forma de los nódulos es importante a la hora de diagnosticar, también lo es la posición, pues un nódulo localizado en el lóbulo superior del pulmón también es considerado peligroso (Wyker, 2021). Esta métrica está normalizada al número total de imágenes que se poseen de cada paciente a lo largo de la altura, por ello es más representativa que la localización vertical del nódulo.

$$\hat{z} = z/H$$

  siendo *H* la altura total del pulmón.

Las siguientes características de los nódulos son propiedades que se utilizan en diversos contextos de CV o reconocimiento de imagen. En este trabajo, intentaremos aplicarlas al problema para probar si son útiles a la hora de clasificar nódulos, lo cual, aunque no esté en la línea metodológica tradicional, podría aportar nuevas perspectivas a la detección del cáncer en imágenes de CT.

- **Histograma de Gradientes (HoG)**: Es una forma de codificar una imagen en función de la orientación del vector de gradientes de cada píxel. El histograma de gradientes es frecuentemente utilizado en detección de objetos, inclusive en la detección de nódulos pulmonares (Jayalaxmi, 2017).
- **Momentos invariante de Hu (Hu)**: Estos calculan a partir de los momentos estadísticos de la máscara binaria y son invariantes a la rotación, escala y traslación. Al igual que los valores propios del tensor de inercia, también se puede ver su uso relacionado a los nódulos pulmonares (Van-Rikxoort, 2014).
- **Patrones Locales Binarios (LBP)**: Son unos descriptores utilizados en diversos problemas de reconocimiento visual que codifican formas en un parche de tres por tres en base a comparar los valores de cada subgrupo con el del subgrupo central. (Pietikäinen, 2015)

Como cada uno de los marcadores anteriormente descritos (HoG, Hu y LBP) está constituido por varios números, cada nódulo tendrá asociado multitud de valores, por ello para comprimir la información que nos aportan, se trabajará únicamente utilizando proyecciones en las primeras componentes principales (PCA), que son las características de textura que se utilizarán en adelante para caracterizar un nódulo

Todos estos descriptores se han extraído usando distintas librerías *Scikit-Image*.

3.2.7 – Agregador por paciente

Llegados a este punto, tenemos gran cantidad de información local de los TACs, los nódulos existentes, su posición, y una serie de características. Sin embargo, las etiquetas de la BDD DSB sólo existen a nivel de paciente, que indican si el paciente fue diagnosticado con cáncer de pulmón o no a un año de la toma de la imagen.

Este es el gran reto en este ejercicio, ya que se debe que determinar si el paciente padece cáncer de pulmón en base a los nódulos detectados, que no están etiquetados.

La forma de obtener métricas a nivel de paciente a partir de las métricas a nivel de nódulos en la herramienta base fue el uso de estadísticos simples, como las características máximas, mínimas o la media de estas a través de todos los nódulos del paciente.

3.2.8 – Clasificador

El último paso de la herramienta es obtener una clasificación binaria a partir de los resultados del agregador por paciente, es decir, de todas las características asociadas a un determinado paciente. Para esto, en el trabajo previo, se usó una GBM (de sus siglas en inglés *Gradient Boosting Machine*). Esta parte del código no se pudo recuperar del trabajo previo, por tanto, se aprovechó para desarrollar un nuevo clasificador. Se ahondará más sobre este en el capítulo 4.

CAPÍTULO IV: Mejoras de la herramienta

A continuación, se describirán las aportaciones de este trabajo a la herramienta base.

4.1 – Selección de características para el entrenamiento del clasificador

Retomando el final del capítulo 3. Los descriptores utilizados en la herramienta de la que parte este trabajo fueron estadísticos que incluían la totalidad de los nódulos (por ejemplo, media y máximo de cada una de las características encontrada). Esta aproximación puede ser insuficiente por distintas razones:

- El uso de estadísticos de promediado o descriptores distribucionales puede causar que las características de un paciente con cáncer de pulmón que también tenga nódulos malignos queden enmascaradas debido a la influencia de los nódulos benignos en las características más relevantes.
- Fijarse únicamente en el valor máximo de las características de un nódulo puede limitarnos en el sentido de que, aunque el valor sea muy alto, no estamos tomando la información del nódulo en conjunto. Por ejemplo, si vemos que un nódulo tiene una forma muy alargada, pero no tenemos tanto en cuenta el tamaño del nódulo, puede que estemos fijándonos en un nódulo muy pequeño, y que incluso ni siquiera sea un nódulo sino un vaso sanguíneo y haya sido mal etiquetado en algún paso anterior. Análogamente, si tomamos como métrica el diámetro máximo, podríamos estar considerando al paciente como afectado de cáncer cuando, si miramos la forma del nódulo, podría quedar claro que es un nódulo benigno, por su redondez o su homogeneidad.

En este trabajo, por tanto, se realizó un estudio paramétrico usando conjuntos extensos de características y combinaciones de ellas, basadas en criterios utilizados por los radiólogos para determinar la posibilidad de cáncer en los pacientes. Para decidir las características finales se estudió, en los distintos árboles entrenados, cuáles fueron las características con mayor capacidad discriminatoria.

Por último, se añaden métricas a nivel de paciente como número de nódulos totales, o número de nódulos que cumplen ciertas características, como una forma de introducir una visión general del paciente huyendo de una selección que se vea sesgada por la presencia de nódulos benignos. También se añadió una nueva característica, alargamiento vertical que es cociente entre Δz y A_{ch} ya que se comprobó que mejoraba sustancialmente el clasificador.

4.2 – Entrenamiento del clasificador

Estos descriptores, ahora de paciente, se utilizaron para alimentar un clasificador de árboles de decisión generado a partir del método XGBoost que es un método reconocido por su efectividad en problemas de clasificación a partir de variables numéricas de las observaciones (en nuestro caso el valor de las características) (nvidia, 2022). El nombre XGBoost viene de *eXtreme Gradient Boosting*, y lo que hace es generar gran cantidad de árboles de decisión iterativamente y evaluar la gradiente entre el árbol generado en el paso anterior y los nuevos, para determinar cual ofrece la mayor disminución de error.

Un árbol de decisión separa secuencialmente a los pacientes aplicando umbrales sucesivos a las distintas características de cada observación (en este caso, cada paciente). Los hiperparámetros más notables son:

- **Profundidad del árbol:** determina cuantas divisiones hay como máximo desde el inicio de la clasificación hasta el resultado final. A mayor profundidad, la complejidad del problema que se puede tratar es mayor, pero también aumentan los tiempos de cálculo (al nivel de tamaño de BDD y características que estamos trabajando esto no es problemático mientras se mantenga un número razonable) y se es más propenso a caer en sobre-entrenamiento, lo que quiere decir que el problema se ha ajustado demasiado a los datos de entrada y no es generalizable, esto supondrá una caída en los resultados con datos nuevos y se debe evitar. Cuando un clasificador generaliza bien, no hay sobre-entrenamiento, se le denomina más conservador.
- **Peso mínimo del hijo:** Define la mínima suma de pesos en todas las observaciones requeridas para un hijo. A mayor valor, modelos más conservadores rango $[0, \infty)$.
- **Muestra de columnas por árbol/nódulo:** Define el número de características entre el total que se utilizarán para hacer cada interacción, impide el uso continuado de las mismas características conforme se baja el valor, lo cual ayuda a generalizar el modelo. El rango permitido es (0,1].
- **Submuestras:** Análogamente al parámetro anterior define el número de muestras (pacientes) que se usan en cada iteración, a menor valor, más general el modelo, pero también mucho más costoso de entrenar. El rango permitido es (0,1].
- **Término de regularización euclídea (λ):** Regularización del término L_2 en los pesos, a mayor valor, modelos más conservadores, por defecto es 1.
- **Término de regularización dispersa (α):** Regularización del término L_1 en los pesos, a mayor valor, modelos más dispersos y por lo tanto más conservadores, por defecto es 0. Acelera el proceso en casos de alta dimensionalidad. La búsqueda de modelos dispersos (*sparse*) se conoce como *sparse regularization* y permite obtener modelos llamados "más parsimoniosos".
- **Ganancia mínima exigida en cada nodo (γ):** Especifica la ganancia mínima en cada nódulo. A mayor gamma más conservador, rango $[0, \infty)$.
- **Reducción de los pesos por iteración (η):** Reducción de los pesos en cada iteración subsecuente, previene sobre-entrenamiento. El rango permitido es (0,1] pero los valores típicos recomendados son [0.01-0.2].
- **N.º de árboles paralelos:** Cuántos árboles se generan en cada iteración. A mayor número de árboles mayor necesidad computacional. Un número demasiado alto puede incurrir a que el modelo no generalice bien por la misma mecánica que tener muestreos muy altos.
- **N.º de rondas:** Determina el número de iteraciones que se harán en total, un número demasiado alto puede provocar sobre-entrenamiento. Este parámetro en realidad no es tan importante porque usaremos un método llamado *early-stopping.*
- ***Early-stopping*:** Para cada iteración (ronda) se reserva un número de observaciones con las que se evaluará la LL de la iteración actual. El método *early-stopping* registra las LL recientes y aborta el entrenamiento anticipadamente cuando estas empiezan a aumentar, debido al sobre-

entrenamiento. El modelo resultante es aquel de la interacción con menos LL.

Además de los parámetros, otra opción que tenemos es el método por utilizar, para el propósito de nuestros algoritmos contamos con dos opciones. El primero de ellos es, *exact*, que es un algoritmo voraz que calcula el gradiente máximo en cada iteración, este es un método que requiere de más recursos computacionales, especialmente a mayor número de características. El segundo, *hist* calcula una aproximación, y por un coste computacional mucho menor adquiere unos resultados muy parecidos.

La Figura 11 ilustra como funcionaría un árbol de decisión. Un hijo es todo lo que hay aguas debajo de una flecha, un nódulo es cada una de las cajas azules.

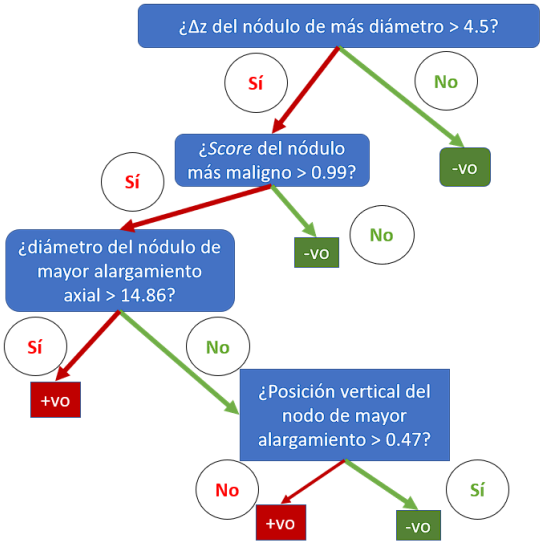


Figura 11. Ejemplo del funcionamiento de un árbol de decisión

Adicionalmente, el XGBoost, es un algoritmo de *boosting*. Este tipo de algoritmos unifican la información de varios modelos más débiles, como en este caso sería un árbol de decisión, para generar un modelo más complejo y robusto, un ejemplo de esto se encuentra en la Figura 12, donde se presentan una serie de árboles con sus ROC y la ROC del modelo que los aglutina.

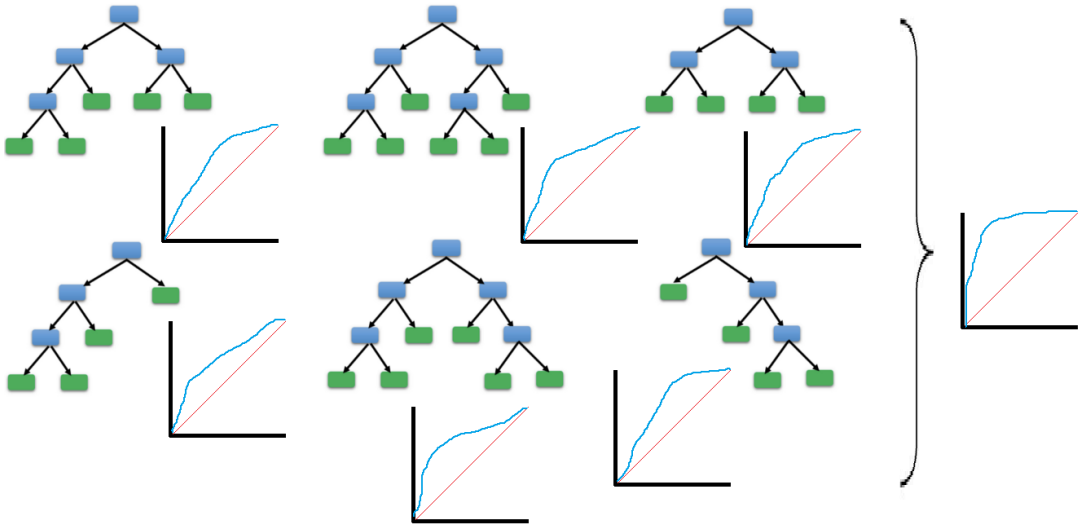


Figura 12. Esquema ilustrativo del funcionamiento del boosting como agregador de clasificadores

Para complementar la elección de características de entrada se estudiaron los parámetros de ajuste del algoritmo para sacar el máximo rendimiento en cada uno de los grupos de características elegidos. Fijándonos en la LL sobre los datos de entrenamiento y validación se eligieron los parámetros como profundidad del árbol, o criterios en base a los cuales añadir nuevas divisiones en este para evitar sobre-entrenamiento, pero también para no tener modelos excesivamente generalizados donde el rendimiento es bajo incluso en los datos de entrenamiento, estos parámetros se recogen en la Tabla 4.

Parámetro	**Valor**
Profundidad	4
Submuestras	0.5
Muestreo por árbol	0.5
Muestreo por hijo	0.5
Peso mínimo del hijo	12
γ	3
λ	2
α	1
η	0.05
N.º de árboles paralelos	120
N.º de rondas	160
Early-stopping	8

Tabla 4. Parámetros finales del clasificador

El entrenamiento del clasificador se hace en dos fases, primero, con un conjunto muy extenso de características se hace un primer entrenamiento, con el método *hist*. Una vez terminado se extraen los datos de la importancia que cada característica ha tenido en la clasificación, que equivale a la ganancia que ha provocado la inclusión de la característica en el entrenamiento total. Típicamente un número reducido de características, alrededor de la decena, tienen una importancia mucho mayor que el resto, y a partir de las dos decenas el resultado del clasificador deja de mejorar. Posteriormente se guardan las características que mayor importancia han tenido en la primera fase y se utilizan en un nuevo modelo que se entrenará esta vez con el método *exact*.

4.3 – Red de malignidad

Durante la DSB, muchos equipos entrenaron "redes de malignidad" a partir de la base de datos LIDC-IDRI para incorporar el conocimiento de los radiólogos sobre la malignidad de los nódulos, y utilizar este conocimiento para determinar el riesgo de cáncer de los pacientes.

Esta información puede sernos relevante, ya que nos estamos fijando en los nódulos para determinar si un paciente tendrá cáncer o no. Sin embargo, un paciente puede tener nódulos malignos a causa del cáncer o benignos, derivados simplemente de una calcificación, o incluso es posible que se haya detectado un artefacto (o un vaso sanguíneo) como nódulo sin que este lo sea. Es importante saber en qué objetos debemos fijarnos para poder diagnosticar al paciente con la mayor precisión posible. Además, una vez clasificado el paciente, nuestro objetivo con esta herramienta es poder justificar nuestra decisión. Un marcador que indique al radiólogo qué nódulos son más probablemente malignos nos puede ayudar en estas situaciones.

El trabajo original propone el uso de una ResNet50 con parches de 40×40×3 píxeles, centradas en los respectivos nódulos. Al no estar el modelo disponible se aprovechará para probar distintas arquitecturas de la red para conseguir un modelo que se ajuste lo mejor posible a los datos disponibles.

Para alimentar la red de malignidad se utiliza un código que, a partir de los archivos XML relacionados con los pacientes de la BBDD LIDC-IDRI que indican donde están los nódulos según la lectura de los médicos extrae las coordenadas de los nódulos y su malignidad. Posteriormente se identifica a que píxeles en los pacientes preprocesados correspondían los marcados por los radiólogos en los ficheros DICOM originales y genera las imágenes del tamaño adecuado para la red, en la Figura 13 se pueden observar distintos nódulos con distintas malignidades.

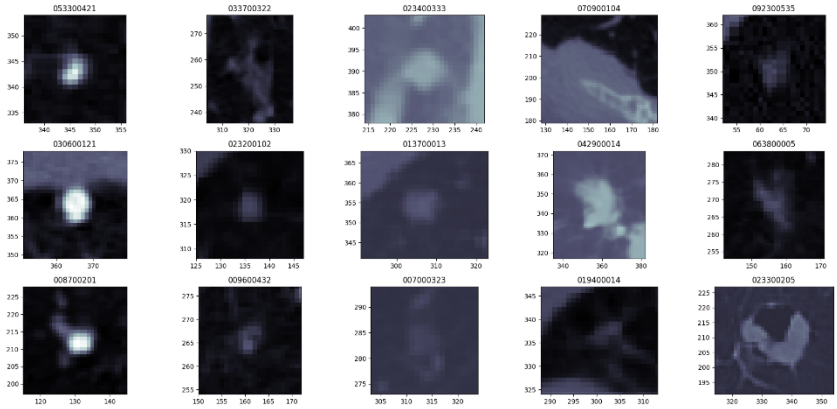


Figura 13. Nódulos anotados por los radiólogos, de menos malignidad (izquierda) a más malignidad (derecha)

Para entrenar la red y conseguir los mejores resultados existen varias opciones:

- Actuar sobre la topología de la red (profundidad, número de capas internas, tamaño de los filtros convolucionales).
- Actuar sobre los hiperparámetros de la red (*learning rate*, balanceo de momentos, paso mínimo etc.)
- Actuar sobre las entradas, distinto tamaño de parches, tanto en lado como en profundidad, compensación de las distintas clases.
- Actuar sobre las salidas, aquí las vías a explorar son: extraer una salida (regresión) y tratar la entrada como un continuo de malignidad, entre 0 y 1; o establecer una salida para cada uno de los valores de malignidad marcados por los radiólogos, del 1 al 5. En este segundo caso será necesario evaluar como traducir el resultado, de cinco regresiones a un único valor. Si tomando la clase con valor máximo o haciendo una suma ponderada, por ejemplo.
- Utilizar *Data Augmentation.* Mientras que el modelo es sensible a cambios en la orientación de los nódulos, o volteos en los ejes, los nódulos y su malignidad no lo son, por tanto, partiendo de una misma imagen, podemos generar nuevos datos reflejándola o rotándola.

La red de malignidad se entrenó inicialmente ejecutando el código heredado, sobre imágenes de 40×40×3 pixeles, obtenidas de la base de datos de LIDC, según las lecturas de los radiólogos.

Una vez obtenidos resultados iniciales, se hicieron una serie de cambios cuyas conclusiones se recogen a continuación:

- **Profundidad de la red**: se probaron tres opciones, 34, 50 y 101 capas de profundidad. Para los casos de 50 y 101 los errores de entrenamiento eran muy bajos mientras que los de validación estaban lejos de esos niveles, lo cual nos podría indicar que la red no generaliza bien y deberíamos probar arquitecturas menos complejas, es decir, menos profundidad. Sin embargo, los resultados de la red de 34 capas, además de no mejorar en validación empeoraban mucho el entrenamiento, lo cual nos indica que no se le permite suficiente complejidad al modelo para captar todos los patrones. Finalmente nos decantamos por una ResNet50
- **Longitud de lado del parche**: inicialmente se tenían parches de 40×40×3, en el trabajo original se proponía añadir más contexto a las imágenes probando parches mayores, pero los resultados no mejoraban. Tras examinar varios nódulos en su contexto, se llegó a la conclusión de que, para nódulos no muy grandes (menos de 40 píxeles de diámetro), había muchos píxeles de entorno que no aportaban información interesante. Mientras que los nódulos muy grandes (mayores de 40 píxeles) a lo mejor se podrían identificar correctamente sólo con la parte central, ya que la mayoría de ellos serían malignos. Por tanto, se decidió probar en la dirección opuesta y se llegó a una dimensión de parche de 32×32×3, a partir de la cuál las disminuciones no ofrecían mejoría.
- **Profundidad del nódulo**: análogamente se consideró añadir más profundidad a los nódulos, pero, de nuevo, la información adicional no daba buenos resultados.
- **Inclusión de objetos no nódulos**: En las anotaciones de los radiólogos, además de nódulos con malignidad, también figuraban objetos no considerados nódulos, de diámetro mayor de 3 cm, pero sin valor de malignidad. Llegado el momento, se consideró si incluir estos no-nódulos podría ayudar a filtrar también los falsos positivos otorgados por la DL2. Los resultados no favorecían esta hipótesis por lo que se acabó descartando.
- **Número de salidas**: Inicialmente, se tomó la malignidad como una única salida, los resultados fueron muy malos. Cambiar el número de salida a 5 (niveles de malignidad existentes en los datos) para hacer un clasificación multiclase dio mucho mejores resultados que una regresión. Para luego poder evaluar la red se tomó la media ponderada de las salidas como valor otorgado.

Hubo dos otras modificaciones que resultaron de ayuda. Por un lado, balancear el *dataset*, ya que, como se puede ver en la Tabla 5, Las clases 1 y 2 están sobrerrepresentadas. De la mano de este cambio, se incrementó el uso de aumento de datos para permitirnos descartar ejemplos de las clases mayoritarias.

Clase	Ocurrencias
0	885
1	1374
2	2158
3	793
4	599

Tabla 5. N.º de nódulos en cada de malignidad según la calificación de los radiólogos

Como se ha descrito, el modelo final para la detección de la malignidad consiste en una ResNet50, cuya entrada es de tamaño 32×32×3 y cuya salida es de cinco niveles. Para entrenar la red se seleccionó el método Adam. El valor de los hiperparámetros utilizados para el entrenamiento pueden verse en la Tabla 6. Se utilizaron un 30% de los datos para validación y 70% para el entrenamiento.

Parámetro	**Valor**
N.º aumentos	32
Batchsize	32
N.º epochs	1000
Muestras por epoch	900
N.º muestras de validación	1000

Tabla 6. Parámetros de la red de malignidad

CAPÍTULO V: Resultados

En este capítulo se describen los resultados obtenidos con la herramienta. Como se ha dicho, el objetivo del trabajo no solo es desarrollar un clasificador que funcione correctamente, sino una herramienta que sea útil para el radiólogo clínico. Para ello, en primer lugar, se ilustrarán y describirán los resultados de cada uno de los módulos por separado, y posteriormente se realizará un análisis global de la herramienta *end-to-end* como clasificador para el diagnóstico de cáncer.

5.1 – Preprocesado

A continuación, se adjunta en la Figura 14 el resultado del preprocesado en dos pacientes distintos, como se puede apreciar si se compara la columna izquierda con la derecha, la región del pulmón ahora ocupa una mayor porción de la imagen a costa de haber recortado los márgenes y parte de las capas más superficiales del paciente. En ambos casos la distancia de píxel era mayor a la normalizada y por ello se ven aumentadas.

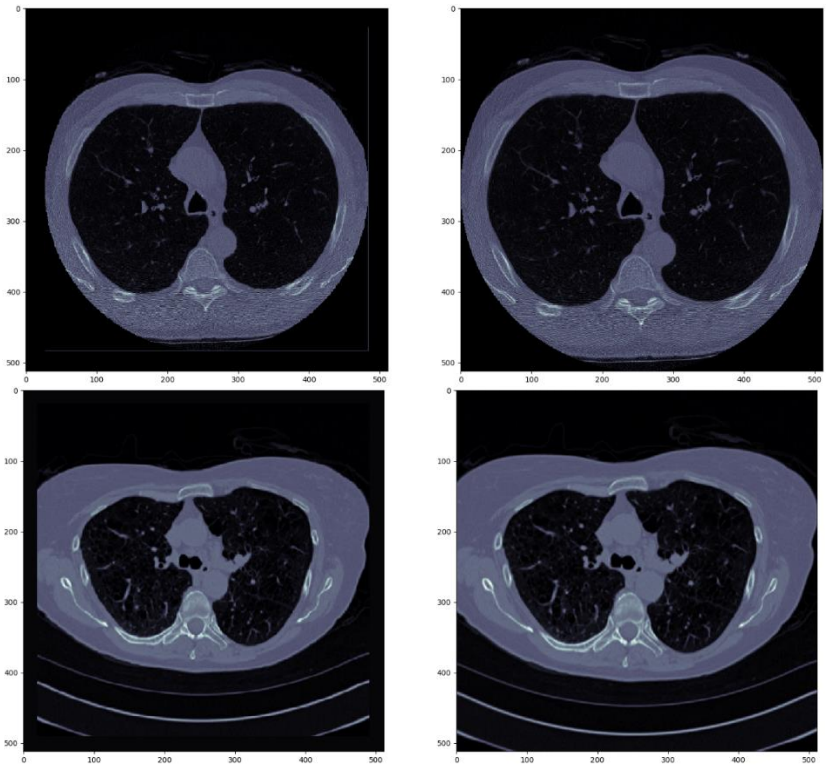


Figura 14. Demostración del preprocesado para dos pacientes, columna de la izquierda, imágenes originales, columna de la derecha, preprocesadas. Un paciente en cada fila.

Seguidamente se evalúa la etapa de segmentación. En la Figura 15, se observa el resultado de la segmentación del pulmón en dos pacientes. Se observa que en uno de los casos el pulmón ha sido segmentado perfectamente, mientras que, en el otro, se ha incluido la región del corazón y los vasos y arterias circundantes (se ha seleccionado un caso en el que la segmentación fracase especialmente). Aunque lo deseable es que suceda como en el primer caso descrito, la segunda segmentación también reduce enormemente la zona de análisis y no perjudica el tratamiento "aguas abajo del proceso" pues el grueso de la zona de no interés se ha marcado correctamente y más importantemente, el pulmón está incluido en su totalidad.

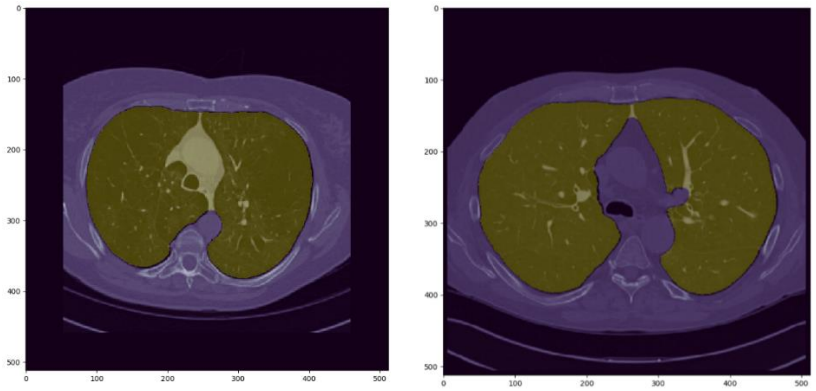


Figura 15. Segmentación de dos pacientes, en amarillo, el área correspondiente al pulmón y en morado el resto. Cabe destacar la diferencia en tratamiento de la zona del corazón y vasos sanguíneos principales. Esta disparidad, mientras que no deseable, tampoco causa limitaciones en el marco de trabajo en el que nos encontramos.

El preprocesado es una tarea completamente automatizada, y el código que la compone trata de solventar diversas problemáticas que puedan ocurrir con los pacientes, sin embargo, esto no siempre es posible. En aquellos casos en los que el código encuentre una excepción no anticipada, el paciente no se preprocesará y la información que contiene no será aprovechada. Esto, a la hora de introducir pacientes nuevos, puede resolverse caso por caso por un técnico, pero no así a procesar una BDD con miles de pacientes. Resultado de esto, la cantidad de datos resultantes se recoge en la Tabla 7:

Base de datos	**Pacientes iniciales**	**Pacientes procesados**	**% de pérdida de TAC**
DSB2017	2058	1571	23.7%
LIDC	1308	976	25.4%

Tabla 7. Número de datos inicial y tras pasar el postproceso para las dos BDD disponibles

Una vez preprocesado el paciente es momento de extraer aquellos elementos de información con los que vamos a trabajar, las ROI. Aplicar el umbral seleccionado en los pacientes obtenemos un resultado como el que se puede ver en la Figura 16. Se delimita la región que circunscribe cada una de las regiones y estas regiones conformarán los datos que se analizarán por el modelo DL1 y seguidamente por el modelo DL2.

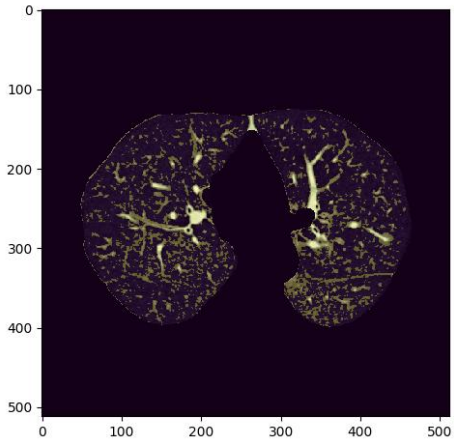


Figura 16. Segmentación del pulmón tras aplicar el umbral que determina las regiones de interés

5.2 – Resultados de los modelos de aprendizaje profundo DL1 y DL2

En la Figura 17 se muestran los *scores* asignados por el modelo DL2 a aquellos nódulos que el modelo DL1 ha considerado de la clase positiva (es decir, nódulos). De entre los

5.545.958 objetos identificados en la BDD DSB2017, 3.766.306 casi el 70% de los señalados tienen puntuaciones inferiores a 0.001. Este resultado es esperado según lo explicado en los apartados 3.2.3 y 3.2.4 dónde se describe el fundamento de estos modelos.

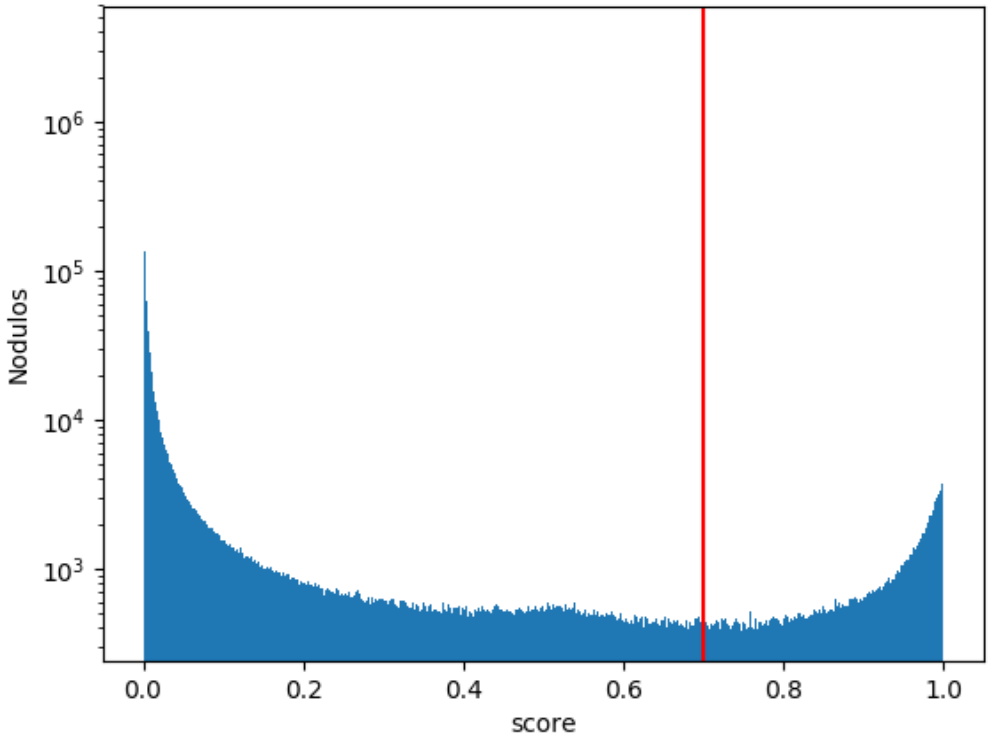


Figura 17. Distribución de los scores de los nódulos detectados con la DL1 tras evaluarlos con la DL2. Línea vertical marca el valor de corte (0.7)

En la Figura 18 hay ejemplos tanto de nódulos con scores muy bajos, como de scores intermedios sin llegar al umbral impuesto de 0.7. Este umbral se ha de elegir de forma que tengamos suficiente confianza en que los objetos indicados son nódulos, pero intentando no descartar nódulos en busca de esa seguridad, ya que, en caso de elegir un umbral demasiado alto, estaríamos reduciendo nuestros datos disponibles para continuar con el análisis descartando información valiosa.

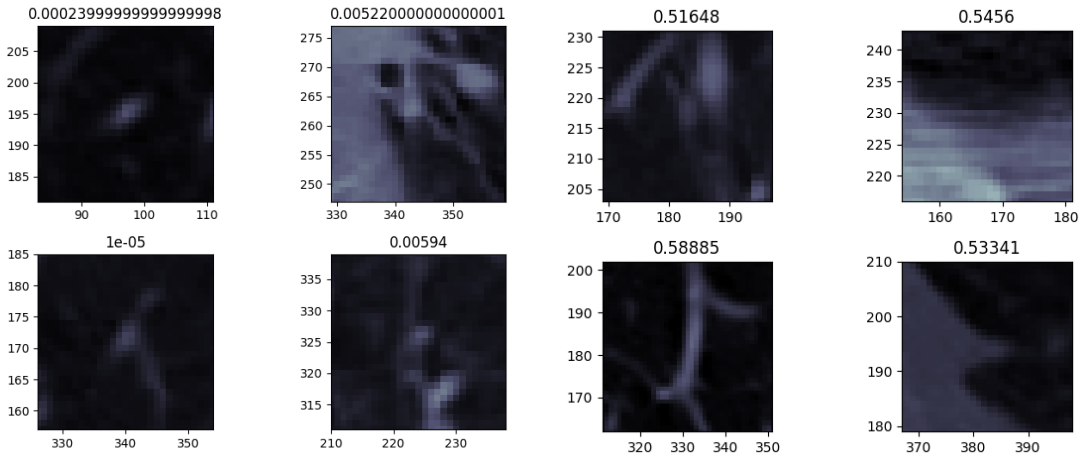


Figura 18. Ejemplos de candidatos a nódulos con scores muy bajos y altos, pero no suficiente como para ser considerados nódulos

5.3 – Características

Una vez han sido identificados los nódulos, y rechazados aquellos cuyo score es menor al 0.7 según la DL2, nos encontramos con un total de 90.580 nódulos, distribuidos entre todos los pacientes. A continuación, se repasarán algunas de las características y se mostrarán nódulos característicos con los valores más altos o bajos de estas.
Consideremos en primer lugar el diámetro de nódulo. Si bien está métrica es importante a la hora de clasificar nódulos generalmente, podemos ver en la Figura 20 que muchos de los nódulos de mayor diámetro presentan características que en la literatura se consideran poco nocivas, como son la homogeneidad del nódulo, su forma regular y los bordes suaves, aunque solo por el tamaño se considerarían peligrosos. Se consideró el filtrar los nódulos más pequeños ya que, puede parecer por las imágenes expuestas que sería difícil obtener buena información de estos, pero los nódulos menores a 5cm, como se puede observar Figura 19 componen la mitad del total, y los resultados finales empeoraban si se descartaban.

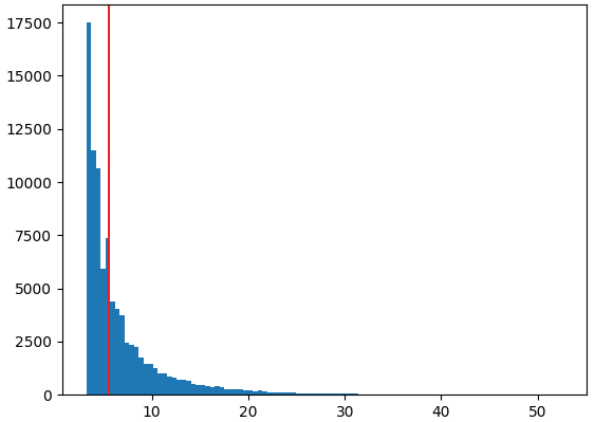


Figura 19. Histograma de diámetros de nódulos

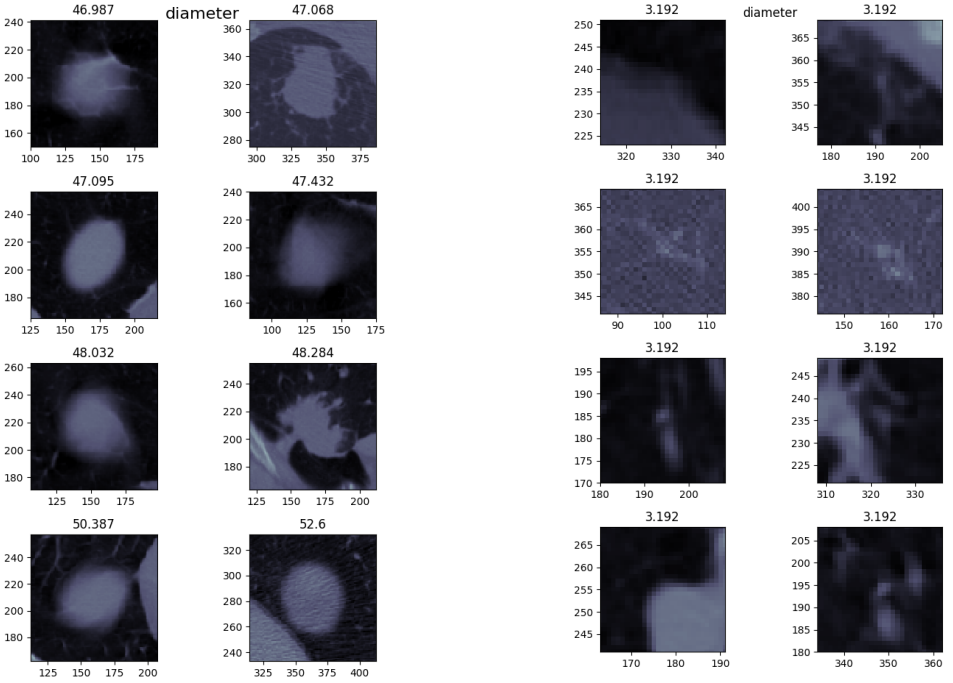


Figura 20. Nódulos de mayor diámetro (columnas de la izquierda) y de menor diámetro (columnas de la derecha)

Seguidamente, tenemos la característica Δz, que es esencialmente la altura del nódulo. Esta característica, a priori, va a ser importante para el clasificador. El problema que existe, a la hora de aplicar las características de alto nivel, es que los valores están muy discretizados y que 82.470 nódulos tienen un valor de 1 en esta característica (más del 90% del total, Figura 21). Entonces no podemos usar el nódulo de máxima altura para clasificar con respecto a este, porque no suele ser único.

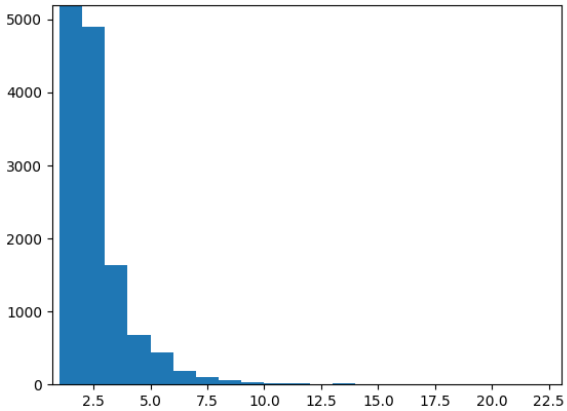


Figura 21. Histograma de altura de los nódulos, el eje de ordenadas se ha cortado en 5000 para mayor claridad del resto de valores

Observando los nódulos de mayor altura en la Figura 22 podemos ver algunos que presentan características de alto peligro, a diferencia de los nódulos de mayor diámetro.

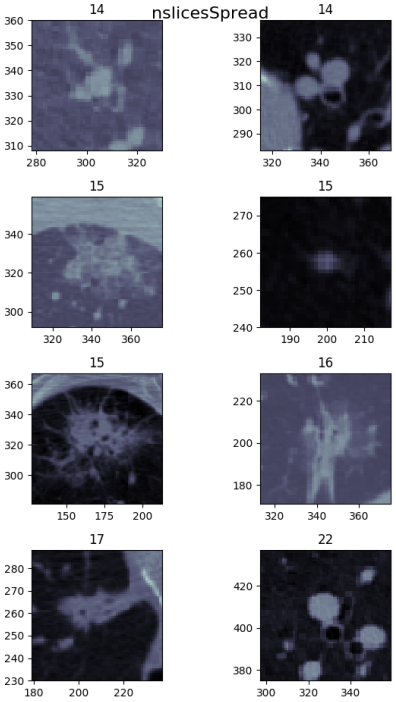


Figura 22. Nódulos de mayor Δz

El score, Figura 23, es una característica interesante en la que fijarse en este apartado ya que no tiene una definición *per se* sobre que implica morfológicamente un *score* más alto o bajo, ya que este valor es determinado por la red neuronal profunda DL2. Parece haber, de todas formas, dos características que propician un *score* alto. En primer lugar, nódulos redondeados y muy brillantes, lo cual no está necesariamente asociado a un nódulo maligno. En segundo lugar, masas muy grandes, irregulares y con ramificaciones, lo cual nos haría sospechar enormemente de estos nódulos si los viésemos en una TAC.

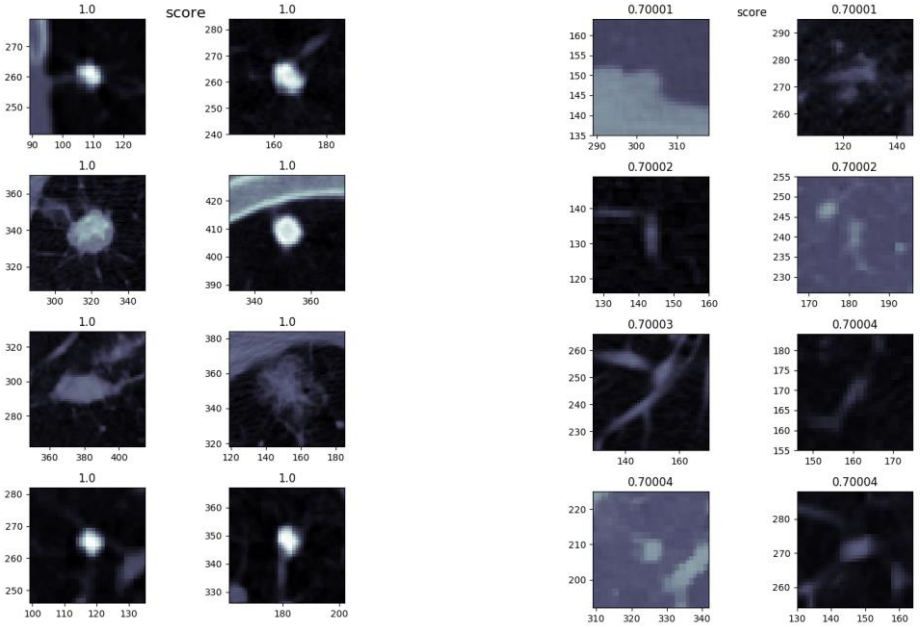


Figura 23. Nódulos de mayor (izquierda) y menor (derecha) score

Similarmente a la Figura 23, en la Figura 24. Nódulos de mayor (izquierda) y menor (derecha) intensidad máxima y media
, que recoge los nódulos con mayores y menores intensidades máxima y media. Encontramos nódulos redondeados entre los de mayor valor, sin embargo, entre los nódulos de menor valor podemos ver alguna forma preocupante (tercera y cuarta fila, sexta columna)

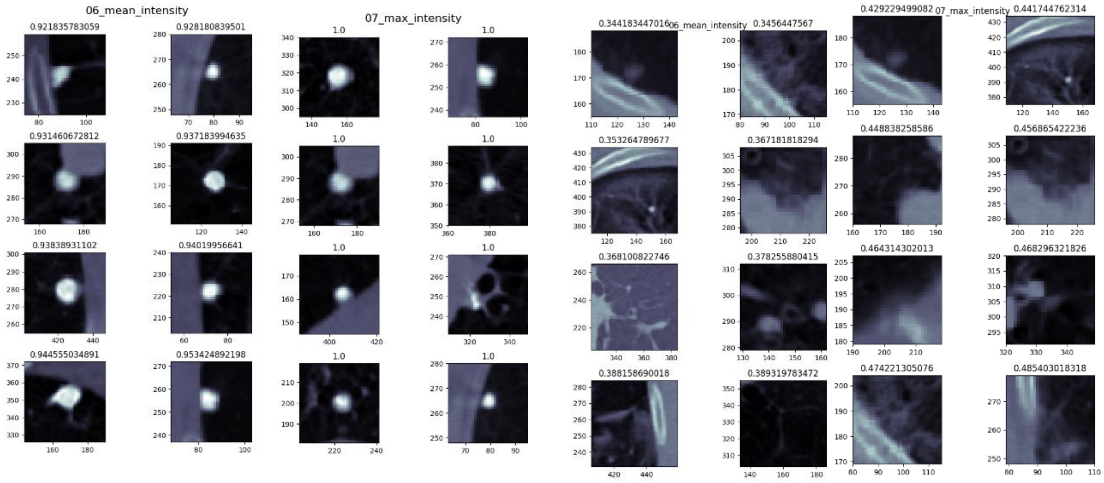


Figura 24. Nódulos de mayor (izquierda) y menor (derecha) intensidad máxima y media

En el caso de la excentricidad, encontramos, como cabe esperar, figuras alargadas, asociadas a mayor excentricidad, como se ve en la Figura 25.

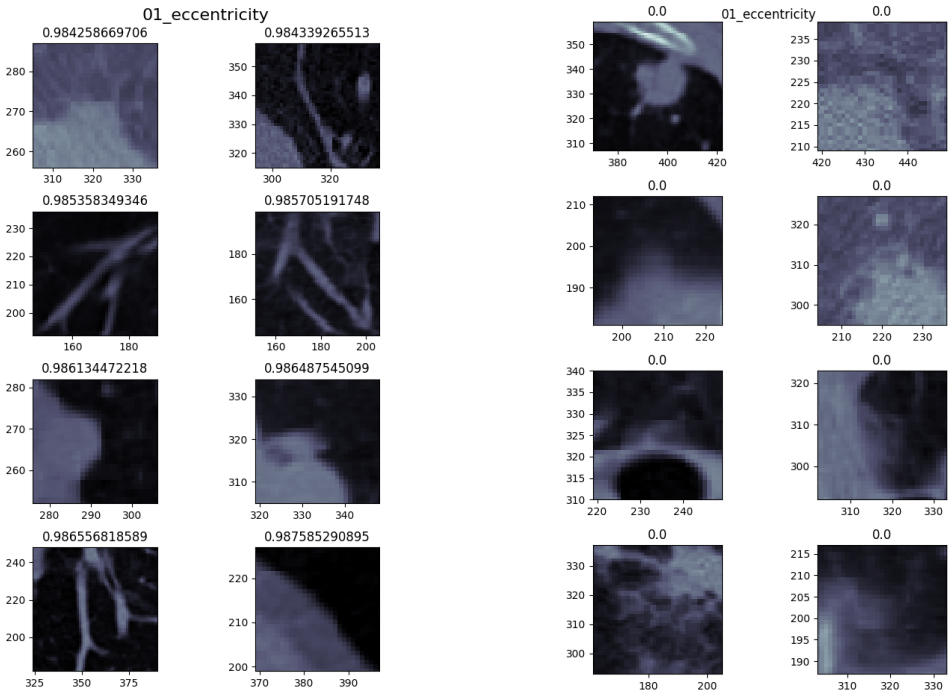


Figura 25. Nódulos de mayor (izquierda) y menor (derecha) excentricidad

Concluyendo respecto a esta característica, los supuestos nódulos más extensos, a la vista de alguien no entrenado en la materia se pueden clasificar en dos grupos, vasos sanguíneos, o nódulos cercanos a la pared del pulmón. En cuanto a los nódulos poco excéntricos hay algunos que pueden parecer llamativos (fila 4, columna 3), pero el problema que encontramos es, que similarmente a los casos anteriores, el valor mínimo se repite mucho, englobando en este grupo nódulos demasiado distintos entre sí como para marcar un criterio basado en esto.

Pasando al área, cuyos nódulos más representativos se pueden ver en la Figura 26 se ven ciertos nódulos preocupantes entre aquellos de mayor valor, aunque no todos. Respecto a los de área menor, el nódulo de la posición (columna 1, fila 4) es algo irregular, pero en el resto es difícil distinguir nada a simple vista, y son en general muy pequeños.

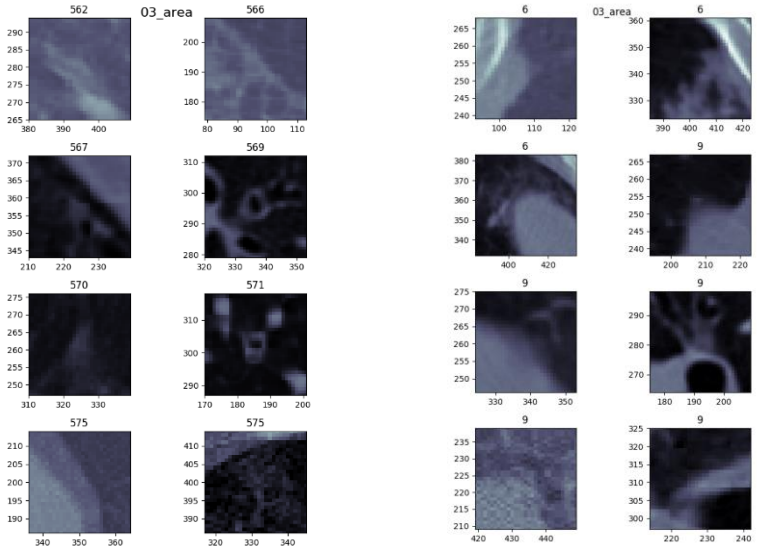


Figura 26. Nódulos de mayor (izquierda) y menor (derecha) área

Respecto al perímetro, la Figura 27 ilustra nódulos muy interesantes, aunque para algunos pueda haber duda de si realmente se tratan de nódulos o de vasos (los dos más inferiores de la imagen izquierda), hay otros que sí parecen ser claramente sospechosos de malignidad. En contraste, los nódulos con pequeño perímetro parecen todos benignos.

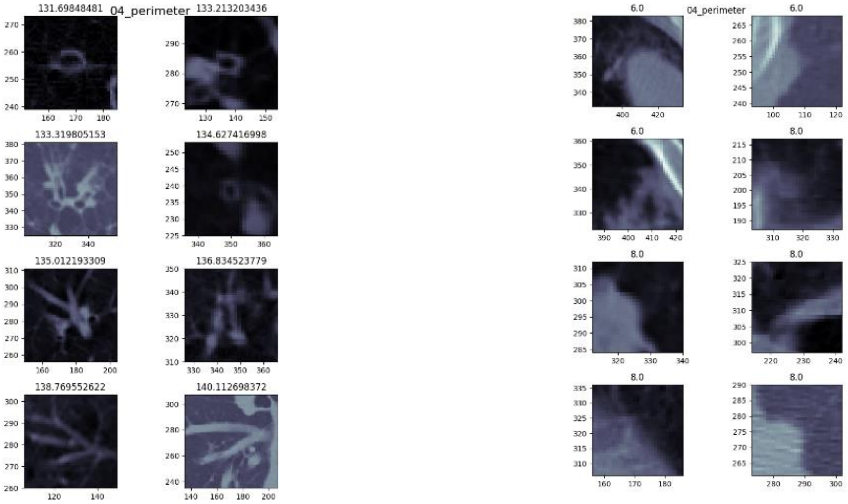


Figura 27. Nódulos de mayor (izquierda) y menor (derecha) perímetro

La Figura 28 muestra en su parte izquierda nódulos que claramente cumplen con las características que se presentan como malignas en la literatura, y los nódulos de poca espicularidad axial, a su vez, son todo lo contrario. Este tipo de patrones son lo que se considerarían muy prometedores a la hora de clasificar.

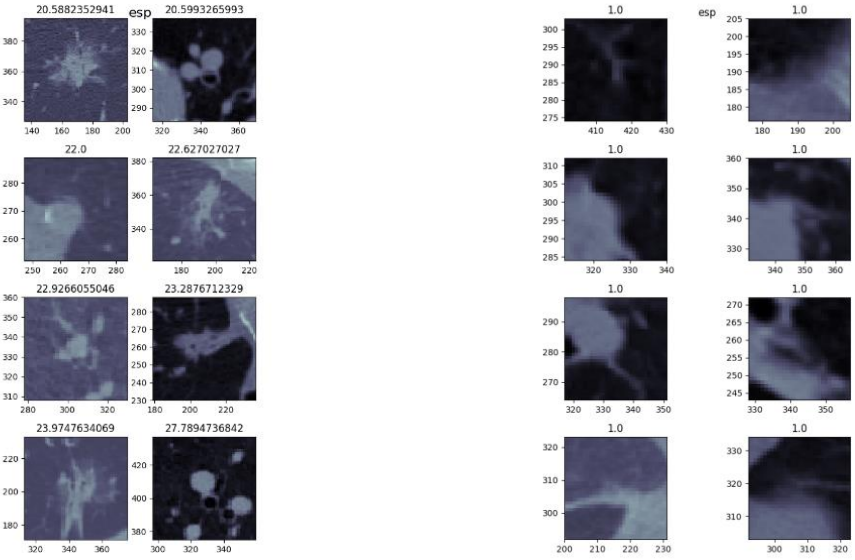


Figura 28. Nódulos de mayor (izquierda) y menor (derecha) espicularidad axial.

A pesar de lo descrito en la bibliografía sobre nódulos presentes en la parte superior de los pulmones, la Figura 29 no parece dejar evidencia de distinciones morfológicas que indiquen claramente una mayor peligrosidad de ninguno de los grupos de nódulos, si bien esta es una característica posicional y no de forma ni textura.

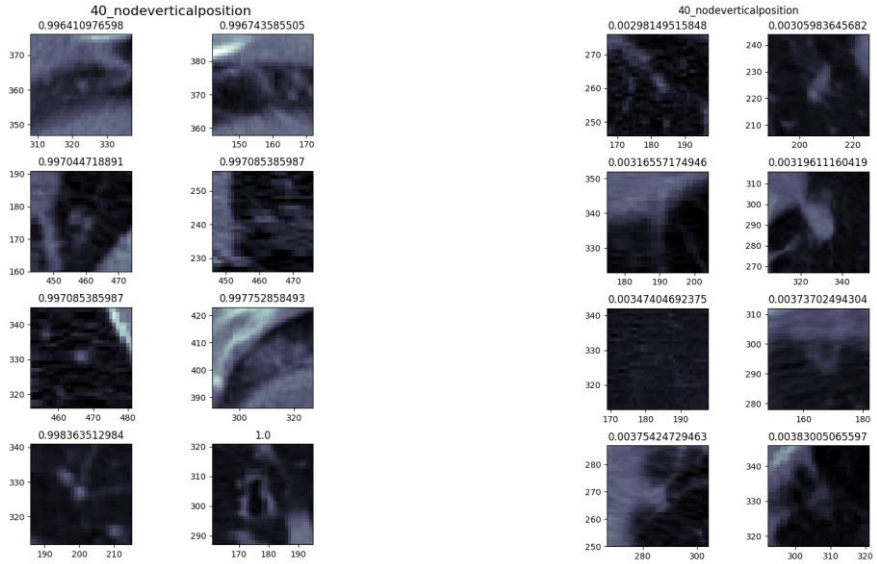


Figura 29.Nódulos de mayor (izquierda) y menor (derecha) posición vertical relativa.

En la Figura 29, Figura 30 y Figura 31 tenemos los nódulos representativos de la solidez, extensión y homogeneidad, estas tres características se han agrupado porque similarmente entre todas ellas, en los nódulos máximos es difícil discernir lo que se observa, aunque si pueda distinguirse un hallazgo que parece preocupante (primera columna segunda fila en la Figura 30 y segunda columna primera fila en la Figura 31). Para los nódulos mínimos se encuentra, como en otras características, que los nódulos representativos son más bien vasos sanguíneos.

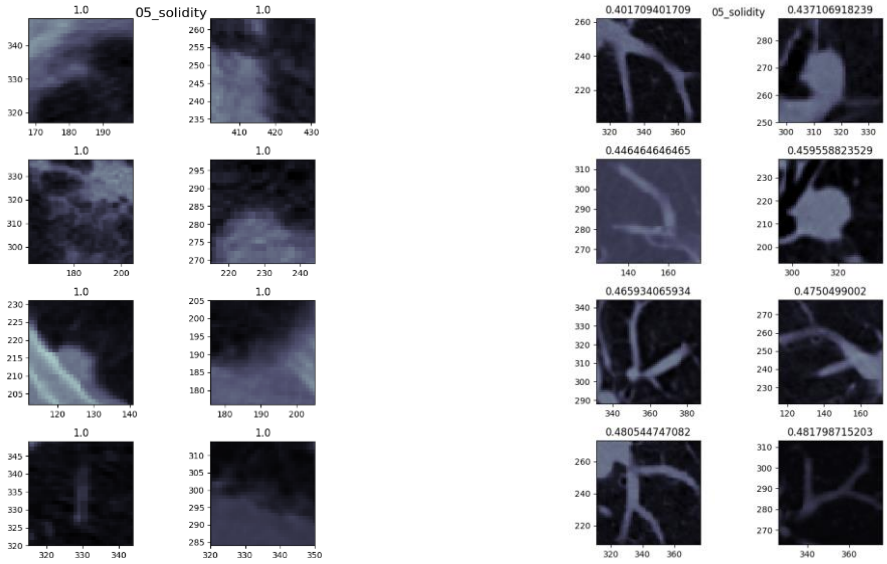


Figura 30. Nódulos de mayor (izquierda) y menor (derecha) solidez

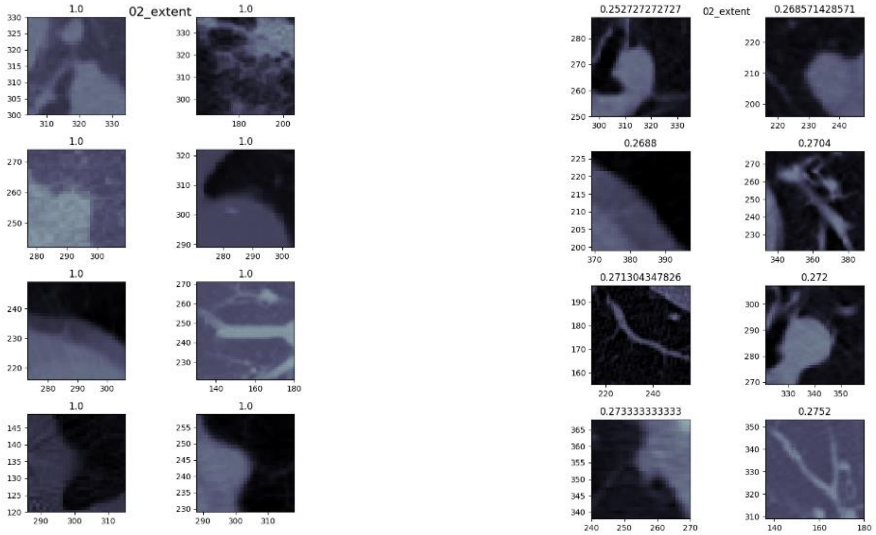


Figura 31. Nódulos de mayor (izquierda) y menor (derecha) extensión

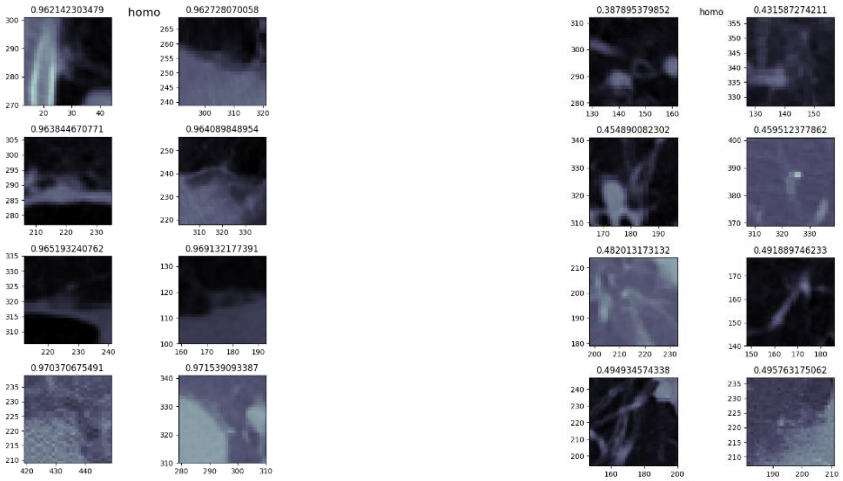


Figura 32. Nódulos de mayor (izquierda) y menor (derecha) homogeneidad

Finalizando este apartado, de la Figura 33 a la Figura 36, tenemos ejemplos de las características morfológicas, HoG, Hu, LBP y autovalores del tensor de inercia. En general los patrones para cada uno de los grupos son claros, prácticamente todos los nódulos de cada grupo guardan similitud, lo cual es una cualidad deseable en un descriptor. Las que parecen albergar más nodos preocupantes son las componentes principales 1 y 3 de los LBP, y los valore extremos en los tensores de inercia. De nuevo, hay muchos de los grupos que, a simple vista parecen falsos positivos, o nodos de pequeña preocupación, regulares, pequeños y brillantes.

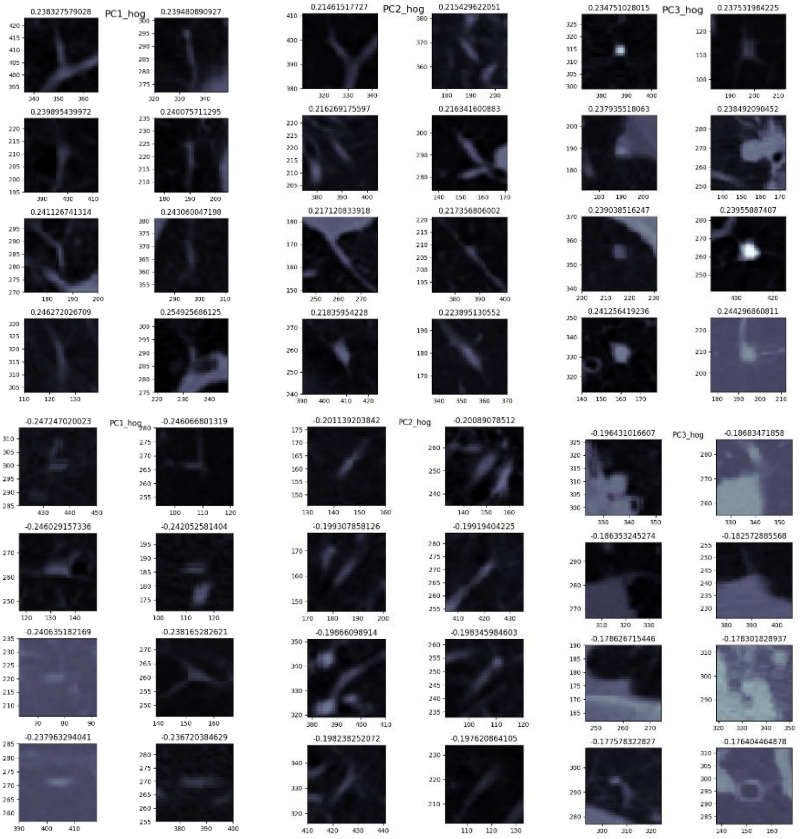


Figura 33. Nódulos de mayor (superior) y menor (inferior) componentes principales del histograma de gradientes

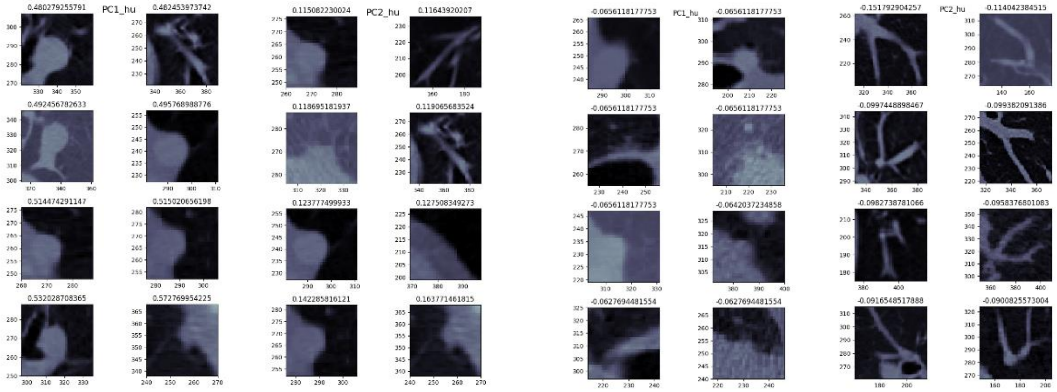
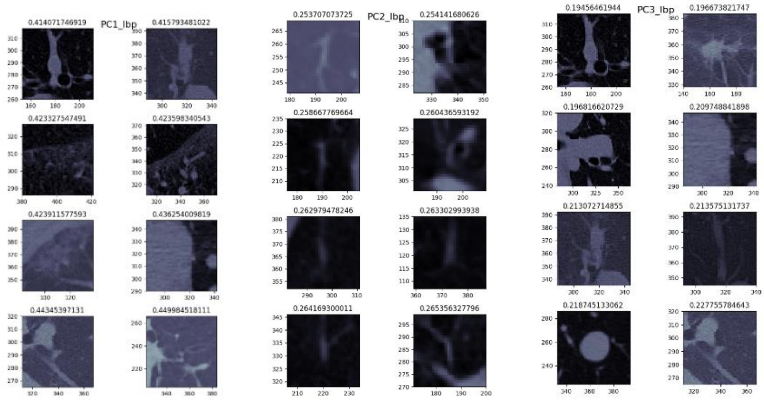


Figura 34. Nódulos de mayor (izquierda) y menor (derecha) componentes principales de invariantes de Hu



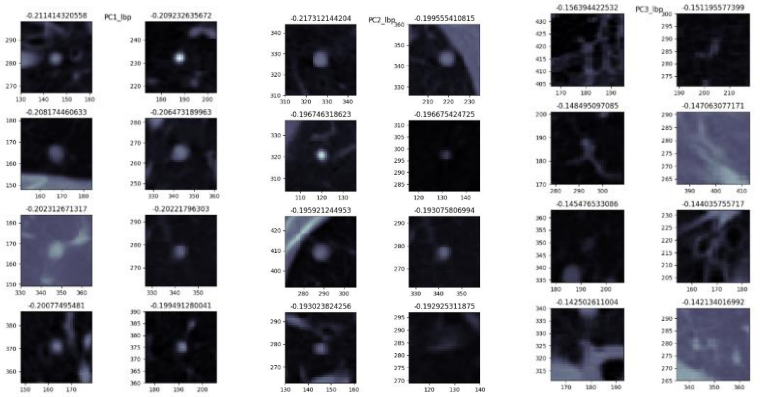


Figura 35 Nódulos de mayor (superior) y menor (inferior) componentes principales de los patrones binarios locales

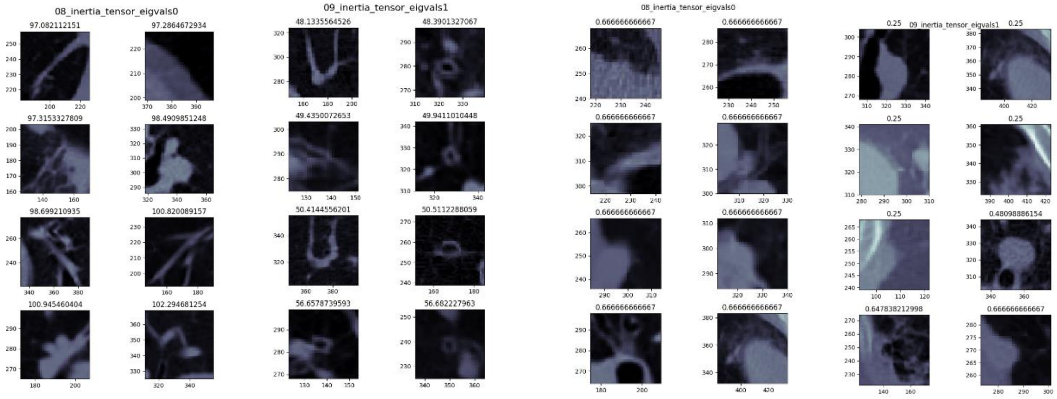


Figura 36. Nódulos de mayor (izquierda) y menor (derecha) valores propios de los tensores de inercia

Es necesario recordar que, a la hora de clasificar, todas estas características se habrán agregado a nivel de paciente de una forma más o menos racional, siguiendo estándares clínicos. Así, todas las características que se han descrito en este apartado se combinan de forma compleja y agregada, mientras que hasta ahora las hemos analizado de forma individual.

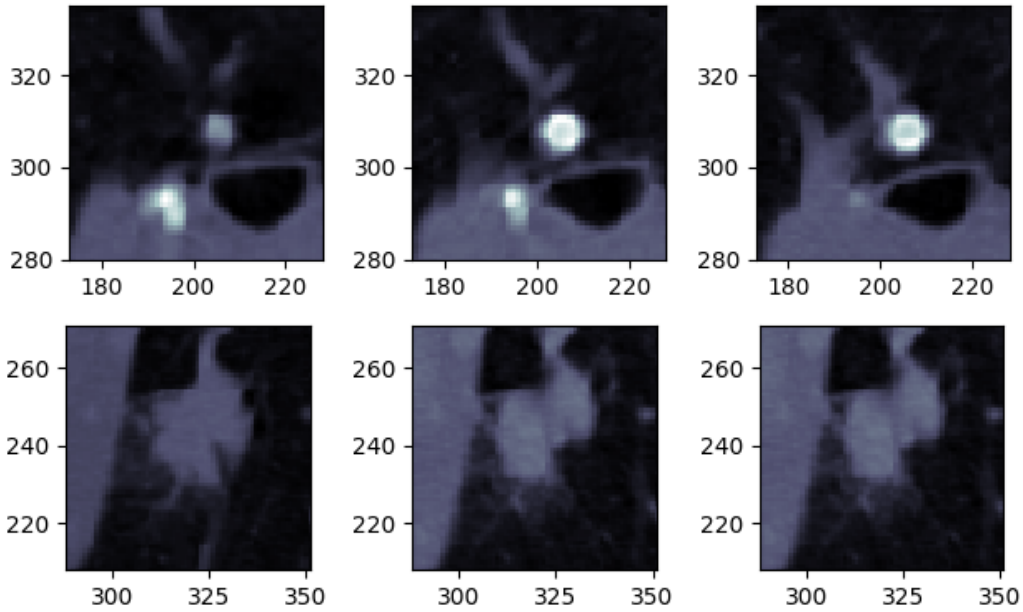


Figura 37. Dos ejemplos de nódulos con varios cortes de cada uno

Para terminar, en la Figura 37, se ilustran dos nódulos diferentes, mediante varios cortes sucesivos, con las consiguientes características asociadas reportadas en la Tabla 8. El primer nódulo tiene claramente las características de un nódulo benigno, siendo el segundo maligno con mayor seguridad.

Característica	**Superior**	**Inferior**
Diámetro	21.76	27.41
Δz	4	7
Score	1.00	0.99
Excentricidad	0.43	0.88
Extensión	0.79	1
Área	123	18
Perímetro	38.97	14
Solidez	0.97	1
Intensidad media	0.82	0.72
Intensidad máxima	1.00	0.87
Alargamiento en z	5.07	7.00
λ_1, λ_2	10.87, 8.88	2.92, 0.667,
PC HoG	-0.03, 0.02, -0.01	0.04, 0.01, 0.02
PC Hu	-0.01, 0.01	-0.01, 0.00
PC LBP	0.19, -0.11,0.06	0.20, -0.07, 0.02

Tabla 8. Características correspondientes a los nódulos de la Figura 28

5.4 – Red de malignidad

Para la red de malignidad primero se extrajeron los nódulos de la BDD LIDC según las indicaciones propuestas por los radiólogos en los ficheros XML y se anotaron los nódulos siguiendo una codificación dependiente del paciente y la malignidad del nódulo hecha de tal forma que fuese única para cada nódulo.

La matriz de confusión final de la red de malignidad se puede observar en la Figura 38, cada celda contiene el valor de ocurrencias de nódulos de un valor real de malignidad

determinado a las que se les ha asignado un valor usando el modelo. Por tanto, los valores en la diagonal son clasificaciones correctas, y conforme nos alejamos de esta son erróneos en mayor medida. Por ejemplo, la celda (fila 1, columna 0) recoge el número de veces (26) que hemos clasificado un nódulo de malignidad 1 como nódulo de malignidad 0, mientras que la celda (fila 4, columna 0) recogería cuantos nódulos de malignidad 4 hemos clasificado como malignidad 0, que sería el caso más grave (5).

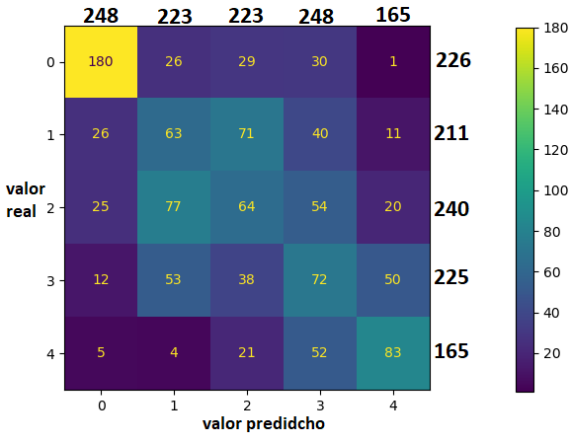


Figura 38. Matriz de confusión red de malignidad

Como cabría esperar, se observa en la Figura 39 y Figura 40 que a medida que aumenta la malignidad de los nódulos otorgada por la ResNet, estos toman características más parecidas a las descritas en la bibliografía como nódulos malignos, si bien no es así para todos los casos. Principalmente, hay nódulos etiquetados como malignos, columna 1 fila 3 de la Figura 40, que a simple vista no parecen muy preocupantes. Lo que sí se puede afirmar es que los nódulos de la Figura 39 sí que parecen ser todos de baja malignidad.

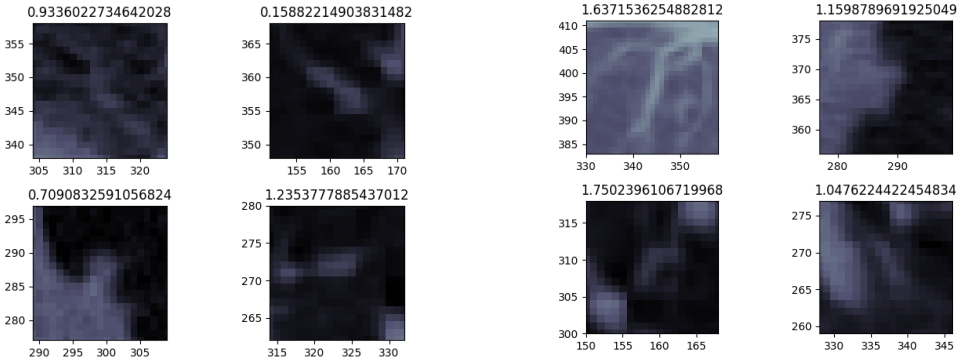


Figura 39. Nódulos de baja malignidad según la ResNet entrenada en este trabajo

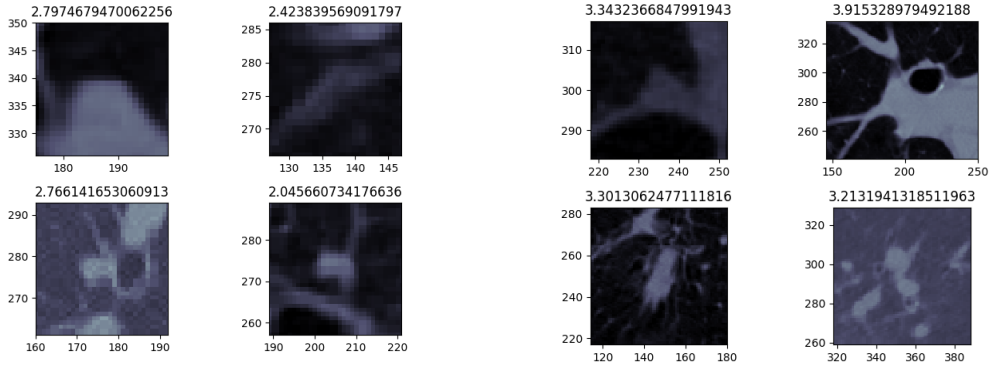


Figura 40. Nódulos de alta malignidad según la ResNet entrenada en este trabajo

En aras de ilustrar los datos sobre nódulos reales, los nódulos de la Figura 37 tienen valores de malignidad de 0.002 (superior) y 3.41 (inferior).

5.5 – Estudio de la relación entre malignidad y otras características

Como forma de intentar arrojar luz sobre los resultados de la red de malignidad y además evaluar el poder discriminativo de las características, se aplicó el método de extracción de características a los nódulos utilizados para generar la red de malignidad y se observó la distribución de cada característica para los nódulos separados por su valor de malignidad.

En primer lugar, en la Figura 41 podemos ver tres características con un claro patrón discriminatorio: el diámetro, la posición relativa y la alargamiento axial. Otro patrón notable, aunque menos claro, es el descenso en el n.º de corte medio para los nódulos más malignos, lo cual es también una característica que se puede encontrar en bibliografía al respecto del diagnóstico desde un enfoque más médico tradicional (Khan, 2011). Una apreciación general que extraemos de este estudio es que la distribución de características para los nódulos menos malignos suele abarcar un rango mayor de valores que las otras clases, aunque por supuesto una afirmación de este tipo requeriría de un estudio más pormenorizado.

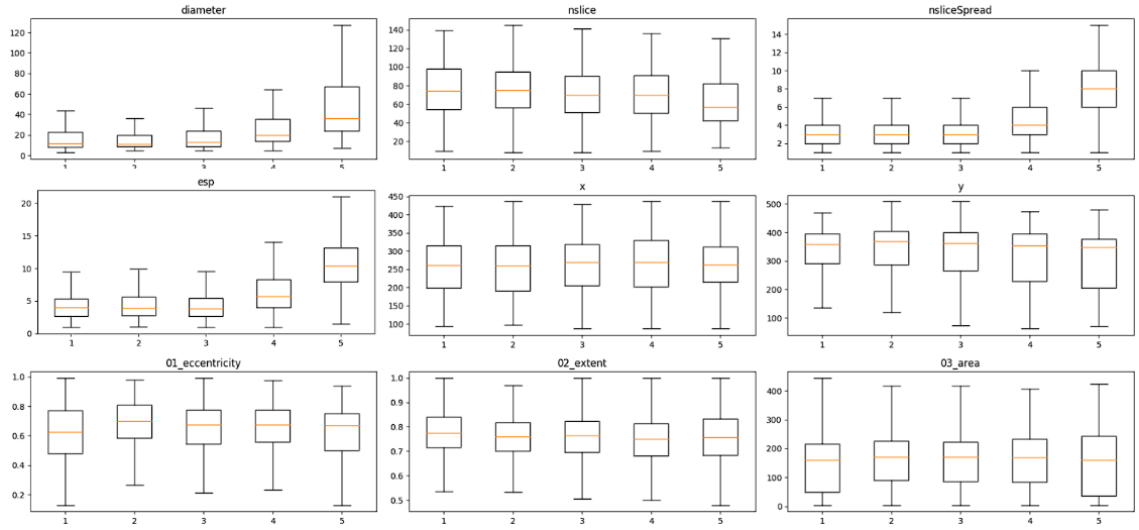


Figura 41. Distribuciones de las características en los nódulos separados por malignidad I. de izquierda a derecha y de arriba abajo: diámetro, n.º corte, extensión en z, alargamiento axial, posiciones x e y, excentricidad, extensión, área.

Continuando con el estudio en la Figura 42 los patrones más destacables son unas solideces e intensidades máximas muy altas todos los nódulos benignos, valores especialmente altos o bajos de intensidad media para nódulos benignos. Tendencias a crecer ligeramente en los tensores de inercia con la malignidad y a descender con los histogramas de gradientes (ver también la Figura 43).

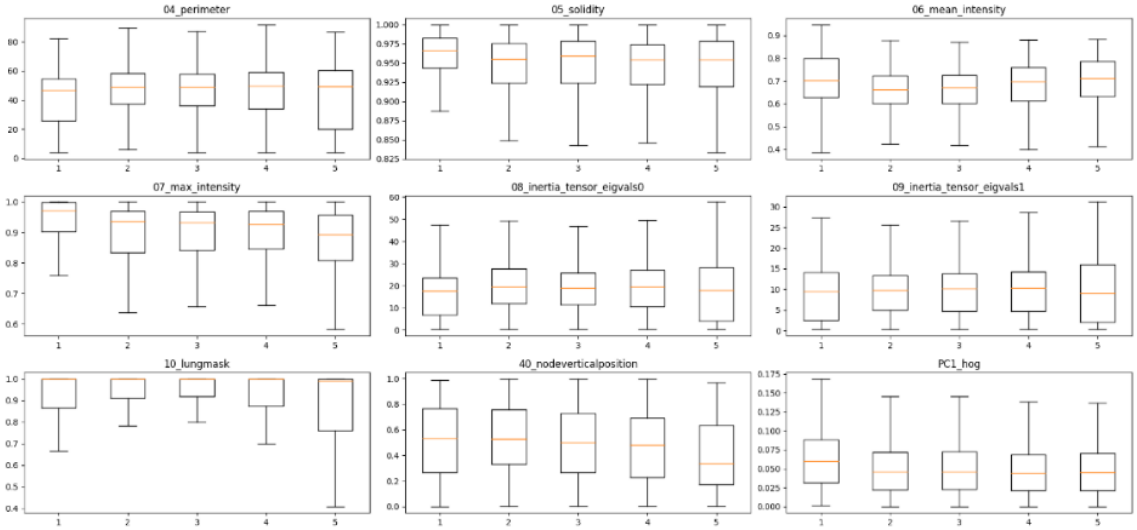


Figura 42. Distribuciones de las características en los nódulos separados por malignidad II. De izquierda a derecha y de arriba abajo: perímetro, solidez, intensidad media, intensidad máxima, tensores de inercia 0 y 1, lungmask, posición vertical relativa,

Para finalizar este apartado, nos centramos en la Figura 43, donde vemos tendencias a valores más bajos de las componentes principales 2 y especialmente 3 del LBP en valores altos de malignidad. Unos valores inesperadamente altos para la homogeneidad de nódulos malignos, una correlación entre la malignidad y la posición más centrada de los nódulos y por último una tendencia de los nódulos benignos a tener valores bajos de la segunda componente principal de los momentos Hu.

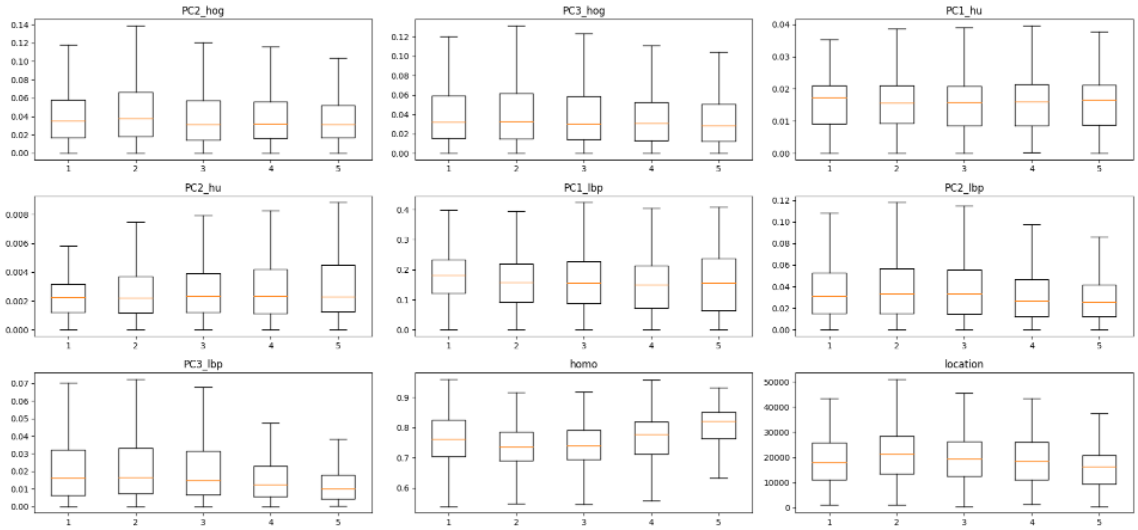


Figura 43. Distribuciones de las características en los nódulos separados por malignidad III. De izquierda a derecha y de arriba abajo: PC2 y 3 HoG, PC 1 y 2 HU, PC 1, 2 y 3 LBP, homogeneidad.

5.6 – Clasificador

Como resultados finales de este trabajo, evaluamos el clasificador final. Estos resultados son los que marcarán el grado de éxito de la herramienta global en su aplicación al proceso de diagnóstico.

Los resultados se presentan en base a tres análisis distintos:

- Uso de descriptores estadísticos globales (mínimos, máximos, medias) frente a descriptores de alto nivel (combinación de descriptores, basada en aspectos clínicos).
- Primera y segunda fase de entrenamiento del clasificador. Descritos en el apartado 4.3.
- Utilización de la malignidad como característica.

Por último, notar que se ha ejecutado un *kfold* de 8 para cada caso, esto quiere decir que los datos se han dividido en 8 grupos iguales y se ha ejecutado el entrenamiento 8 veces, cada vez utilizando un grupo distinto como grupo de validación, así podemos obtener métricas de la robustez del modelo, a mayor varianza de la ROC, menor robustez, esta se puede ver en el sombreado de las curvas de la Figura 44.

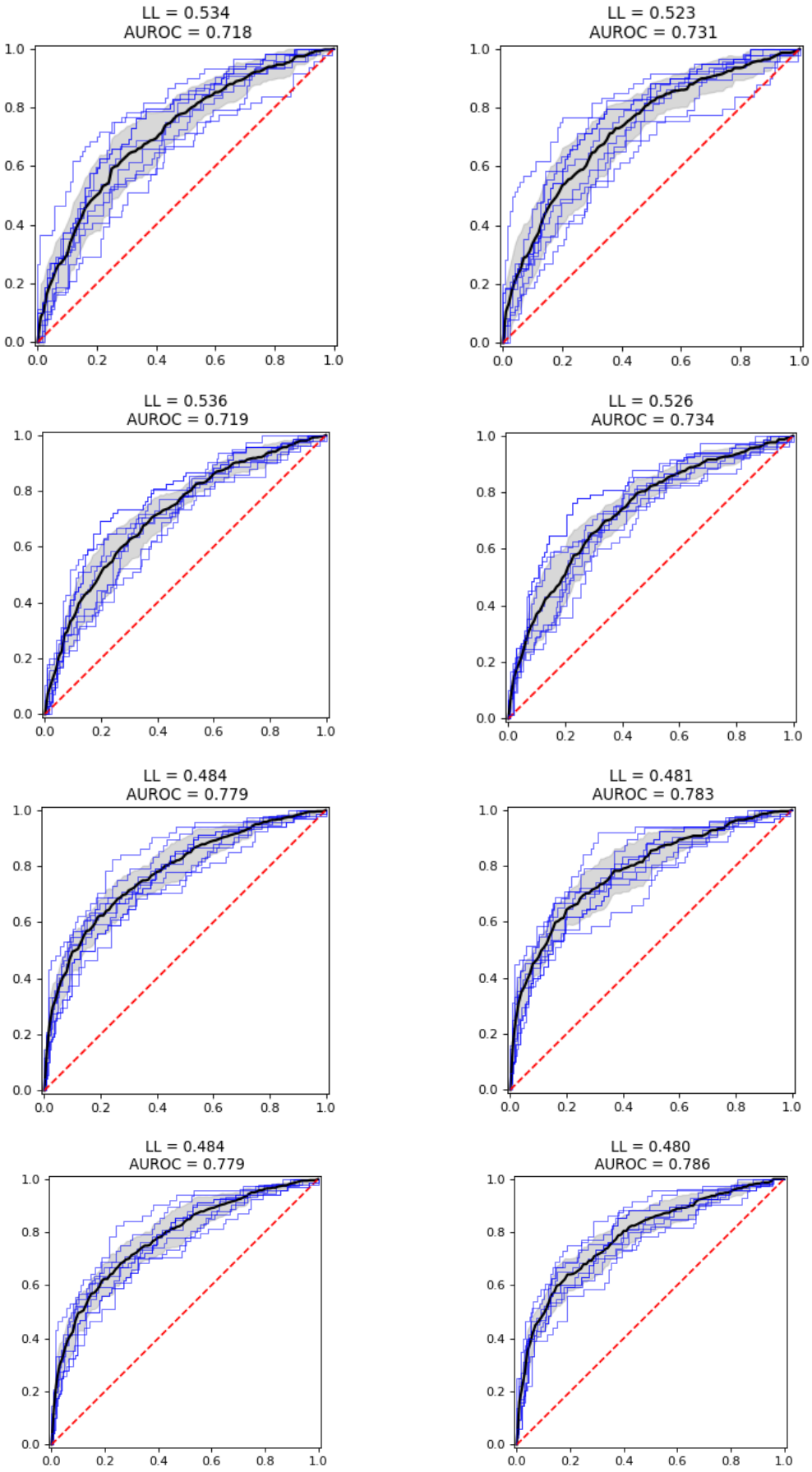


Figura 44. Clasificadores, a la izquierda, primera fase, a la derecha, segunda fase. Filas, de arriba abajo: Descriptores estadísticos sin malignidad, con malignidad, combinados sin malignidad, con malignidad.

Método	Malignidad	Fase	LL	AUC
Estadísticos	No	Primera	0.534	0.718±0.07
		Segunda	0.526	0.731±0.07
	Sí	Primera	0.536	0.719±0.05
		Segunda	0.526	0.734±0.05
Combinados	No	Primera	0.484	0.779±0.05
		Segunda	0.481	0.783±0.05
	Sí	Primera	0.484	0.779±0.05
		Segunda	0.480	0.786±0.05

Tabla 9. Resumen de las LL y AUC en función del método, fase, y utilización de malignidad. AUC presentadas como valor medio a través de los kfold ± desviación típica.

El primer resultado a destacar, como se ve resumido en la Tabla 9, es que el salto en la reducción de LL al aplicar el método de descriptores combinados es de aproximadamente 0.05, dependiendo de en qué método nos fijemos, y el aumento en la AUC de más de 0.05. Esta diferencia de resultados es muy considerable. Para el caso de la malignidad, aunque su impacto positivo no se vea muy reflejado en la LL, sí que hay aumentos de 0.003 en la AUC, que es una mejora bastante reducida, si bien, en la Figura 44 se puede apreciar una menor dispersión a través de los *kfolds* para ambos métodos, al aplicar la malignidad, esto significa que el modelo es más robusto, lo cual es positivo. Adicionalmente, observando las importancias de las características, la malignidad o descriptores relacionados con ella siempre aparecen entre los puestos más altos, si bien nunca los más altos, pues los relacionados con el alargamiento axial de los nódulos dominan a los demás.

Una recopilación de las AUC de los clasificadores finales de los trabajos revisados, junto a las diferentes cifras encontradas para radiólogos, el clasificador inicial y final de este trabajo, y a la máxima y mínima posible, se encuentra en la Figura 45.

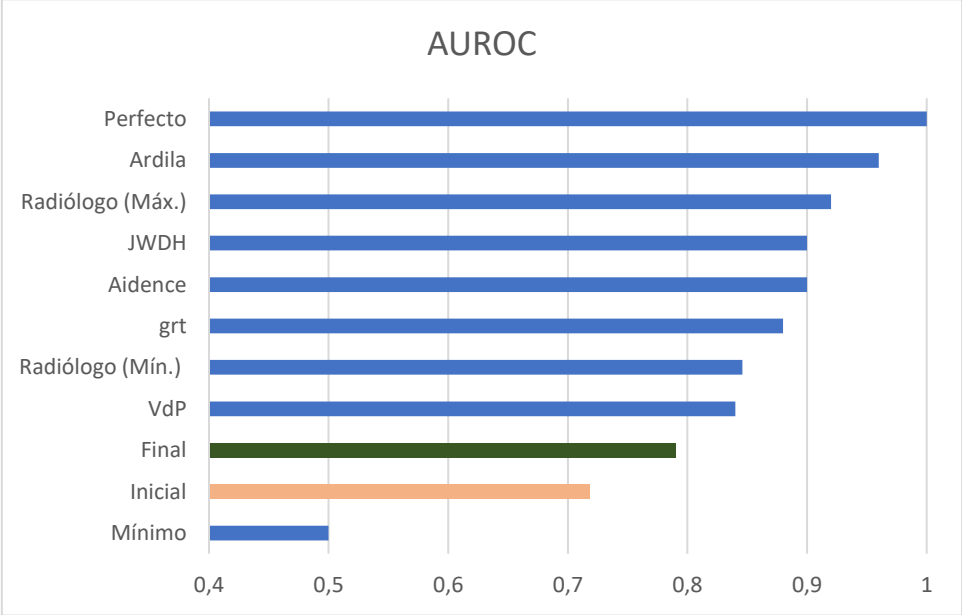


Figura 45. recopilación de las AUC de los clasificadores finales de los trabajos revisados, junto a las diferentes cifras encontradas para radiólogos, el clasificador inicial y final de este trabajo, y al máximo y mínimo posible

CAPÍTULO VI: Conclusiones

6.1 – Resumen del trabajo

En este trabajo se ha creado una herramienta que, a partir del TAC de un paciente, detecta sus nódulos, analiza las características de estos y da un diagnóstico sobre si padece o no cáncer. A diferencia de otros algoritmos de clasificación, de caja negra, que también cumplimentan a esta tarea, la herramienta generada en este trabajo se ha diseñado en base a la transparencia. Los nodos detectados, sus características, y los conjuntos de nódulos y características que más relevantes han sido para el diagnóstico, quedan a disposición del médico. De esta forma, el diagnóstico es explicable tanto para el médico como el paciente. Por último, la idea de uso de esta herramienta es que facilitar el trabajo del médico, no sustituirlo.

6.2 - Principales conclusiones

Este trabajo ha sido concluido con éxito en dos aspectos. Por un lado, la herramienta desarrollada cumple la función de dar un diagnóstico del paciente incluyendo un conjunto de nódulos notables para esta decisión (al margen de la bondad de este). Por otro lado, los dos apartados (red de malignidad y clasificador) que se han desarrollado en este trabajo presentan mejoría respecto a las versiones iniciales de la herramienta.

Creemos que, aunque los resultados no destaquen frente al estado del arte, se hace mucha reflexión sobre las características y se mantiene una filosofía de transparencia que puede ayudar al radiólogo a mejorar su rendimiento y ahorrar parte de su tiempo, en lugar de ser una aproximación sustitutiva.

6.3 – Líneas futuras

A lo largo del trabajo, han surgido ideas cuya implementación quedaba fuera del alcance de este, o cuya implementación excedería su marco temporal, estas se citan a continuación:

- Efectuar un revisado de las partes iniciales de la herramienta para evitar pérdida de datos en el preprocesado. Seguramente esta pérdida de datos (alrededor del 25%) ha causado la disminución del rendimiento de la herramienta respecto de la que se usó en el concurso internacional.
- Añadir algún nuevo método de aprendizaje profundo que detecte objetos que aparecen recurrentemente, como vasos sanguíneos, para que estos no supongan un lastre a la hora de clasificar.
- Aumentar los datos utilizados agregando la BDD de NLST, una red de datos no abierta al público pero que se puede solicitar formalmente.
- Explorar más características o combinaciones de ellas que puedan ayudar a separar pacientes según su estado patológico. Para ello, contar con la opinión de un radiólogo que pueda indicar la malignidad de los nódulos según estas.
- Características morfológicas que codifiquen los nodos tridimensionalmente, más que en el corte.

Para incrementar la usabilidad de esta herramienta a nivel clínico, sería necesario una interfaz gráfica de usuario que guíe al radiólogo por las distintas fases de la herramienta. Para la ejecución de esta parte sería conveniente contar con radiólogos que puedan dar su visión a nivel usuario.

Si se considera importante contar con una buena predicción en la herramienta, visto que recientemente han aparecido artículos con muy buenas características, podría desdoblarse el modelo y predecir con modelos como los del estado del arte, pero manteniendo la parte de extracción de features y combinar el resultado final, sacando así el veredicto según los modelos de redes neuronales convolucionales y las características examinadas en este trabajo. Sería importante mantener la transparencia del modelo predictivo, para lo cual se puede añadir una prueba de oclusión que señale la región que ha causado la clasificación para el caso positivo.

Finalmente, se podrían convertir este trabajo o trabajos similares en proyectos multidisciplinares, porque hubiese sido muy interesante trabajar junto a un radiólogo que pueda dar su opinión sobre lo que se ve en los nódulos, y que alguien con formación en ML pueda traducir las ideas del médico en clasificadores de TAC que funcionen bien.

Referencias

ACS. (12 de Enero de 2022). *cancer.org*. Obtenido de https://www.cancer.org/es/cancer/cancer-de-pulmon/acerca/estadisticas-clave.html

Álamo-Junquera. (2011). Effect of false-positive results on reattendance at breast cancer screening programmes in Spain. *European Journal of Public Health*.

Ardakani. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. *Computers in Biology and Medicine*.

Ardila. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*.

Baade, P. D. (2006). Non-Cancer Mortality among People Diagnosed with Cancer (Australia). *Cancer Causes & Control*.

CDC. (2021). *Centers for Disease Control and Prevention*. Obtenido de https://www.cdc.gov/cancer/lung/basic_info/screening.htm

CEC. (2021). *Contra el Cancer España*. Obtenido de https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-pulmon

Collins. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). *Circulation*.

de Koning, M. e. (2020). Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *the New England Journal of Medicine*.

Dembla, G. (31 de 10 de 2022). *Towards Data Science*. Obtenido de https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a

Ding, Y. (2022). *researchgate*. Obtenido de https://www.researchgate.net/figure/The-architecture-of-Unet_fig2_334287825

Gruy, F. (2015). Inertia tensor as morphological descriptor for aggregation dynamics. .

Gu, Y. (2019). Automatic lung nodule detection using multi-scale dot nodule-enhancement filter and weighted support vector machines in chest computed tomography.

Hammer, S. C. (2020). Factors Influencing the False Positive Rate in CT Lung Cancer Screening.

Henschke. (2006). Survival of patients with stage I lung cancer detected on CT screening. *The New England Journal of Medicine*.

Hillis. (2018). Radiologist performance in the detection of lung cancer using CT. *Epub*.

Houssein, E. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*.

Jacobs. (2021). Deep Learning for Lung Cancer Detection on Screening CT Scans: Results of a Large-Scale Public Competition and an Observer Study with 11 Radiologists. *Radiology: Artificial Intelligence*.

Jayalaxmi. (2017). Classification of Lung Nodules with Feature Extraction using CT scan Images.

kaggle. (2022). *Kaggle Data Science Bowl 2017 Description*. Obtenido de https://www.kaggle.com/competitions/data-science-bowl-2017/overview/description

Kandhasamy, J. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science*, 45-51.

Khan. (2011). Solitary pulmonary nodule: A diagnostic algorithm in the light of current imaging technique. *Avicenna Journal of Medicine*.

Konstantinos Loverdos, A. F. (2019). Lung nodules: A comprehensive review on current approach and management. *Annals of Torathic Medicine*.

Lu. (2013). Suicide and suicide attempt after a cancer diagnosis among young individuals. *Annals of Oncology*.
Macura, A. C. (2012). Evaluation of solitary pulmonary nodule detected during computed tomography examination. *Polish Joutnal of Radiology*.
Ming-Kuei. (1962). *Visual Pattern Recognition by Moment Invariants.*
Moyer. (2013). Screening for Lung Cancer: U.S. Preventive Services Task Force. *annals of internal medicine.*
neumomadrid. (2021). *neumomadrid cribado cancer de pulmon*. Obtenido de https://www.neumomadrid.org/cribado-de-cancer-de-pulmon/
NIBIB. (2022). *www.nibib.nih.gov*. Obtenido de https://www.nibib.nih.gov/espanol/temas-cientificos/tomograf%C3%ADa-computarizada-tc
NIH. (2022). *Instituto Nacional del Cáncer*. Obtenido de https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/metastasis
Noguchi, S. (2022). Deep learning-based algorithm improved radiologists' performance in bone metastases detection on CT. *European Radiology*.
nvidia. (2022). *https://www.nvidia.com*. Obtenido de https://www.nvidia.com/en-us/glossary/data-science/xgboost/
OMS. (21 de 11 de 2022). *https://gco.iarc.fr/*. Obtenido de https://gco.iarc.fr/today/online-analysis-pie?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=total&sex=0&cancer=39&type=1&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=7&group
Patel. (2021). *Cancer.Net*. Obtenido de https://www.cancer.net/es/desplazarse-por-atenci%C3%B3n-del-c%C3%A1ncer/conceptos-b%C3%A1sicos-sobre-el-c%C3%A1ncer/%C2%BFqu%C3%A9-es-la-met%C3%A1stasis#:~:text=El%20c%C3%A1ncer%20de%20pulm%C3%B3n%20tiende,al%20h%C3%ADgado%20y%20los%20pulmones.
Philipp. (2018). *Workshop track - ICLR 2018.*
Pietikäinen. (2015). *Advances in Independent Component Analysis and Learning Machines: Chapter 9 - Two decades of local binary patterns: A survey.* Academic Press.
Reddy, S. (2021). Explainability and artificial intelligence in medicine. *Lancet Digit Healt*.
Remon, D. J. (18 de Diciembre de 2019). *SEOM*. Obtenido de https://seom.org/info-sobre-el-cancer/cancer-de-pulmon
Rudin. (2021). Small-cell lung cancer. *Nature reviews*.
Sanchez-Salcedo. (2015). Lung cancer screening: fourteen year experience of the Pamplona early detection program (P-IELCAP). *Archivos de Bronconeumología*.
SEOM. (2021). *Las cifras del cáncer en España 2021*.
Shakeel. (2019). Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. *Journal of the International Measurement Confederation*.
Shaughnessy. (2017). High False-Positive Rate with Lung Cancer Screening. *Am Fam Physician*.

U.S. department of Health and Human Services. (2022). *www.nibib.nih.gov*. Obtenido de https://www.nibib.nih.gov/espanol/temas-cientificos/tomograf%C3%ADa-computarizada-tc

Van-Rikxoort. (2014). Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images.

Warjurkar. (2021). A Study on Brain Tumor and Parkinson's Disease Diagnosis and Detection using Deep Learning. *Atlantis Press*, 356–364.

WHO. (2021). *World Health Organisation Cancer*. Obtenido de https://www.who.int/news-room/fact-sheets/detail/cancer

Wiener. (2011). Population-Based Risk for Complications After Transthoracic Needle Lung Biopsy of a Pulmonary Nodule: An Analysis of Discharge Records. *Annals of Internal Medicine*.

Wyker. (2021). Solitary Pulmonary Nodule.

Wyman, O. (2020). *El impacto económico y social del cáncer en España.* Asociación Española Contra el Cáncer (AECC).