REPRESENTACIÓN PROFUNDA CON MÚLTIPLES MECANISMOS DE ATENCIÓN PARA LA LOCALIZACIÓN DE NÓDULOS PULMONARES EN SECUENCIAS DE TOMOGRAFÍA COMPUTARIZADA

DAVID DUEÑAS TORRES
ANDRÉS FELIPE MEJÍA PERDOMO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

REPRESENTACIÓN PROFUNDA CON MÚLTIPLES MECANISMOS DE ATENCIÓN PARA LA LOCALIZACIÓN DE NÓDULOS PULMONARES EN SECUENCIAS DE TOMOGRAFÍA COMPUTARIZADA

DAVID DUEÑAS TORRES
ANDRÉS FELIPE MEJíA PERDOMO

Trabajo de grado para optar al título de ingeniero de sistemas

Director:
Fabio Martínez Carrillo
Doctor en Ingeniería de Sistemas y Computación

Codirector:
Luis Carlos Guayacán Chaparro
Magíster en Matemática Aplicada

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

AGRADECIMIENTOS

Primero, a Dios, por permitirme culminar esta carrera y por su guía constante. A mis padres, Abelardo Dueñas y Rosalbina Torres, cuyo apoyo incondicional acompañó cada aprendizaje y a quienes debo profundamente este logro. A todas las personas con quienes colaboré, aun pensando distinto: en la diferencia también se aprende y se avanza. Aprendí a esforzarme incluso cuando los resultados no son los esperados; del fracaso nacen caminos para mejorar, aprender y valorar el proceso.

A mi compañero de grupo, Andrés Mejía, a quien estimo y reconozco como pieza clave en este proyecto. Su disciplina, laboriosidad y compromiso elevaron la calidad del trabajo y hicieron posible cada entrega. Que queden guardados los momentos de risas y esparcimiento como prueba de que, aun en la rigidez académica, se pueden construir ambientes amistosos y colaborativos.

Al grupo de investigación BIVL²ab, por acogerme y orientarme. Al profesor Fabio Martínez, director del grupo, por su guía y respaldo en cada etapa.

Mi gratitud especial al codirector Luis Carlos Guayacán por su entrega, rigor y paciencia, por las horas dedicadas a corregir, enseñar y acompañar este proyecto. Que Dios lo siga bendiciendo para que continúe guiando a más estudiantes por el camino de la ciencia, con la responsabilidad y el carisma que siempre lo han caracterizado.

Por último, y no menos importante, al personal de la universidad, por su excelente servicio durante este proceso y por abrirme las puertas a nuevas experiencias.

David Dueñas Torres

AGRADECIMIENTOS

Expreso mi más profundo agradecimiento a mi director, el profesor Fabio Martínez, por brindarme la oportunidad de ser parte de este grupo de investigación. Agradezco cada espacio de su tiempo, así como la guía necesaria para culminar con éxito este trabajo y el conocimiento adquirido.

A mi codirector, Luis Carlos Guayacán, a quien agradezco por abrirme las puertas al *Detection Team*. No solo ha tenido un papel importante en el desarrollo académico de este proyecto, sino que también ha hecho de la investigación una experiencia disfrutable y tranquila.

A mi compañero, David Dueñas, por compartir momentos en los que la retroalimentación, la discusión y el intercambio de conocimiento fueron parte fundamental de este proceso, así como aquellos fuera del teclado, donde nunca faltaron las risas y la buena compañía.

A mi madre Jacqueline Perdomo y a mi padre Ricardo Mejía, quienes son el motor de mi vida y me han dado el apoyo incondicional en mis capacidades y en momentos difíciles, que desde siempre y por siempre tendrán, y del que yo siempre honraré y estaré agradecido día a día.

A mis compañeros del grupo de investigación BIVL²ab, con quienes no solo compartimos entre todos nuestros trabajos presentados en cada seminario, de los cuales las preguntas y sugerencias fueron clave para enriquecer nuestro aprendizaje, sino también momentos agradables que disfruté desde el primer día de mi llegada y que hicieron de este espacio un lugar de verdadera convivencia y una divertida integración.

Finalmente, agradezco a mis profesores, compañeros, a mi psicóloga y al personal académico y de bienestar, cuyo apoyo y enseñanza fueron un granito de arena que ayudo tanto a mi formación profesional como personal durante estos años.

Andrés Felipe Mejía Perdomo

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

RESUMEN

TÍTULO: Representación profunda con múltiples mecanismos de atención para la localización de nódulos pulmonares en secuencias de tomografia computarizada. [*]

AUTORES: David Dueñas Torres, Andrés Felipe Mejía Perdomo [**]

PALABRAS CLAVE: Cáncer de pulmón, nódulos pulmonares, tomografía computarizada, modelos fundacionales, localización, redes neuronales, mecanismos de atención.

DESCRIPCIÓN: El cáncer de pulmón es la principal causa de muerte por cáncer a nivel mundial (1.8 millones de fallecimientos en el año 2022). Los nódulos pulmonares (NP) son masas potencialmente malignas que constituyen la principal alerta de cáncer. El diagnóstico del cáncer pulmonar se fundamenta en la correcta y temprana localización de los NP en imágenes de tomografía computarizada (TC). Este análisis típicamente se realiza por expertos, quienes armonizando información clínica, factores de riesgo y tomando en cuenta variables de espacialidad observan a plenitud el parénquima para apoyar la tarea de localización. Sin embargo, este procedimiento es subjetivo y los métodos de soporte diagnóstico priorizan patrones locales sin tener en cuenta el contexto espacial del parénquima, relaciones con otras estructuras pulmonares e información complementaria que da soporte al diagnóstico. En este trabajo se implementaron dos arquitecturas para la detección de NP, aprovechando relaciones espaciales de largo alcance en imágenes radiológicas. Se implementaron dos arquitecturas para la detección de nódulos pulmonares (NP). La primera, *RT-DETR*, empleó mecanismos de atención multiescala para mejorar la representación compacta de los NP. La segunda, un modelo fundacional preentrenado (*Grounding DINO*), integró información visual y texto clínico para enriquecer las predicciones. Posteriormente, se aplicó un esquema de destilación *Profesor–Estudiante*, transfiriendo conocimiento del modelo fundacional al compacto. Ambos modelos fueron entrenados y validados en LIDC y NLST. En el conjunto LIDC, el modelo fundacional y el modelo compacto alcanzaron desempeños de CPM de 0.476 y 0.469, respectivamente, mientras que la destilación permitió mejorar el *RT-DETR*, favoreciendo su capacidad de generalización en la localización de NP sobre el NLST.

---

[*] Trabajo de investigación

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez, PhD. Codirector: Luis Carlos Guayacán , Ph.D(c).

ABSTRACT

TITLE: Deep representation with multiple attention mechanisms for pulmonary nodule localization in computed tomography sequences. [*]

AUTHORS: David Dueñas Torres, Andrés Felipe Mejía [**]

KEYWORDS: Lung cancer, lung nodules, computed tomography, foundation models, localization, neural networks, attention mechanisms.

DESCRIPTION:

Lung cancer is the leading cause of cancer-related mortality worldwide (1.8 million deaths in 2022). Pulmonary nodules (PNs) are potentially malignant masses that represent the main early warning sign of cancer. The diagnosis of lung cancer relies on the accurate and early localization of PNs in computed tomography (CT) images. This analysis is typically carried out by experts who, combining clinical information, risk factors, and spatial variables, thoroughly examine the lung parenchyma to support the localization task. However, this process is subjective, and current computer-aided diagnostic methods tend to prioritize local patterns while overlooking the broader spatial context of the parenchyma, its relationships with other pulmonary structures, and complementary information that supports diagnosis. In this work, two architectures were implemented for PN detection, leveraging long-range spatial relationships in radiological images. The first, RT-DETR, employed multiscale attention mechanisms to enhance the compact representation of PNs. The second, a pretrained foundation model (Grounding DINO), integrated visual information and clinical text to enrich predictions. Subsequently, a Teacher–Student distillation scheme was applied, transferring knowledge from the foundation model to the compact model. Both models were trained and validated on LIDC and NLST. On the LIDC dataset, the foundation model and the compact model achieved CPM performances of 0.476 and 0.469, respectively, while distillation further improved RT-DETR, enhancing its generalization capacity for PN localization on the NLST dataset.

---

[*] Research work

[**] Faculty of Physics-Mechanics Engineering. School of Systems Engineering and Informatics. Advisor: Fabio Martínez Carrillo, PhD. Co-advisors: Luis Carlos Guayacán , Ph.D(c).

INTRODUCCIÓN

El cáncer de pulmón (CP) es la principal causa de mortalidad oncológica a nivel mundial, con 2.5 millones de casos nuevos y 1.8 millones de muertes reportadas [1]. Su diagnóstico depende de la detección temprana de nódulos pulmonares (NP), lesiones anómalas que pueden ser malignas y que se identifican mediante tomografía computarizada (TC). No obstante, su pequeño tamaño (entre 3 mm y 30 mm de diámetro) dificulta su detección en estadios iniciales, ya que pueden representar apenas entre el 0.03 % y el 0.3 % del área total de una imagen de TC [2]. Además, sus similitudes morfológicas con estructuras anatómicas, como los vasos sanguíneos, contribuyen a una tasa de omisión de hasta el 25 % en estudios radiológicos [3]. La interpretación visual de las imágenes sigue siendo el principal método de detección, lo que la hace subjetiva y propensa a errores [2,4].

Para abordar estos desafíos, recientemente se han desarrollado estrategias computacionales, principalmente basadas en redes convolucionales (CNN, por sus siglas en inglés), aprendiendo filtros (usualmente de 3×3 pixeles) que extraen progresivamente caracte-

---

[1] W. H. ORGANIZATION. *Global cancer burden growing, amidst mounting need for services*. News release. World Health Organization. Feb. de 2024. URL: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services.

[2] Geoffrey D. RUBIN. «Lung Nodule and Cancer Detection in Computed Tomography Screening». En: *Journal of Thoracic Imaging* 30.2 (2015), págs. 130-138.

[3] Narjust DUMA; Rafael SANTANA-DAVILA y Julian R. MOLINA. «Non–Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment». En: *Mayo Clinic Proceedings* 94.8 (2019), págs. 1623-1640. DOI: 10.1016/j.mayocp.2019.01.013.

[4] Annemilia Del CIELLO, *et al.* «Missed lung cancer: when, where, and why?» En: *Diagnostic and Interventional Radiology* 23.2 (2017), págs. 118-126. DOI: 10.5152/dir.2016.16187.

rísticas con distintos niveles de abstracción semántica en cada capa [5,6,7]. Sin embargo, al estar centradas en regiones locales, estas arquitecturas pueden limitar su capacidad para capturar el contexto global presente en las imágenes de TC, lo que afecta la comprensión estructural completa del parénquima. Alternativamente, los modelos basados en mecanismos de atención han mostrado mayor capacidad para capturar relaciones de largo alcance entre características de la imagen, mejorando la detección de NP [8,9]. Estos modelos dividen la imagen en pequeños parches, los cuales son codificados en vectores y combinados con información posicional para conservar la estructura espacial. Luego, se utilizan mecanismos de auto-atención, que permiten que cada parche se relacione con todos los demás, capturando así información global desde las primeras capas. Esta capacidad de modelar relaciones a larga distancia en la imagen permite al modelo construir representaciones más ricas y contextuales, especialmente útiles para tareas de detección. No obstante, estos métodos, típicamente, requieren grandes volúmenes de datos durante su entrenamiento para lograr un buen desempeño y pueden sobre-ajustarse a conjuntos

---

5 J. DING, *et al.* «Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks». En: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2017, págs. 559-567.

6 Y. SIM, *et al.* «Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs». En: *Radiology* 294.1 (2020), págs. 199-209. DOI: 10.1148/radiol.2019191193.

7 Haichao CAO, *et al.* «A Two-Stage Convolutional Neural Networks for Lung Nodule Detection». En: *IEEE Journal of Biomedical and Health Informatics* (ene. de 2020). DOI: 10.1109/jbhi.2019.2963720.

8 Alexey DOSOVITSKIY, *et al.* «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: *arXiv preprint arXiv:2010.11929* (2020).

9 H. MKINDU; L. WU e Y. ZHAO. «Lung nodule detection in chest CT images based on vision transformer network with Bayesian optimization». En: *Biomedical Signal Processing and Control* 85 (2023), pág. 104866. DOI: 10.1016/j.bspc.2023.104866.

específicos [10]. A pesar de los avances notables evidenciados, la detección de nódulos en TC sigue siendo un problema abierto en la comunidad, con múltiples desafíos en cuanto al modelado de la alta variabilidad en las observaciones, la inclusión de otras fuentes de información así como también la implementación y evaluación sobre bases de datos diferentes, para evidenciar su carácter de generalización.

En este trabajo de grado se implementaron y exploraron dos arquitecturas para la detección de NP, fundamentadas en múltiples mecanismos de atención, que aprovechan tanto relaciones espaciales, como información radiológica complementaria. Por una parte, se adaptó una arquitectura RT-DETR, de tipo *transformer*, que permite una representación de múltiples escalas para la localización y caracterización de objetos de interés compactos *i.e.,* los NP. Por otra parte, se exploró una arquitectura fundacional (Grounding DINO), la cual no solo usa múltiples mecanismos de atención visual, sino que además permite la integración de otras observaciones radiológicas, codificadas como hallazgos radiológicos textuales. Además de ello, esta arquitectura permite aprovechar el contexto de pre-entrenamiento sobre grandes cantidades de datos en imágenes naturales. En una exploración adicional, considerando las características de las dos arquitecturas, se modeló un esquema de aprendizaje Profesor-Estudiante, para destilar información desde el modelo fundacional al modelo RT-DETR. Estos modelos y el esquema de aprendizaje fueron validados en dos conjuntos de datos públicos de imágenes de TC, que además contenían anotaciones sobre la morfología de los nódulos.

---

[10] J. MEI, *et al.* «SANet: A Slice-Aware Network for Pulmonary Nodule Detection». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.8 (2021), págs. 4374-4387. DOI: 10.1109/TPAMI.2021.3058744.

1. FUNDAMENTOS Y TRABAJOS PREVIOS

1.1. CÁNCER DE PULMÓN Y LOS NÓDULOS PULMONARES

El cáncer de pulmón (CP) se sitúa como una de las principales causas de muerte por cáncer en el mundo. En 2022, se registraron alrededor de 2.5 millones de nuevos casos y 1.8 millones de muertes causadas por esta enfermedad [1]. La detección del cáncer de pulmón comienza revisando el historial médico del paciente, un examen físico y la evaluación de imágenes radiológicas o tomografía computarizada (TC). Estas imágenes son capaces de revelar la presencia de masas anormales llamadas nódulos pulmonares (NP), las cuales constituyen el principal indicador de CP. No obstante, los NP son estructuras pequeñas, con diámetros que oscilan entre 3 y 30 mm, lo que representa menos del 0.3 % del tamaño total de una imagen y aproximadamente el 0.013 % del volumen de una TC [2]. Además, presentan una alta variabilidad en su densidad, la cual puede dificultar su identificación. La densidad textural del tejido nodular, reflejada en el nivel de atenuación en la imagen, permite clasificarlos en tres categorías: sólidos, en vidrio esmerilado y subsólidos o parcialmente sólidos. Los nódulos sólidos (los más frecuentes) se caracterizan por una densidad homogénea y baja atenuación. Los nódulos en vidrio esmerilado muestran un aumento de atenuación sin ocultar las estructuras subyacentes, como los vasos sanguíneos. Por su parte, los nódulos sub-sólidos, aunque menos comunes, presentan una mezcla de atenuaciones y son considerados más propensos a ser malignos, estando relacionados con estados precancerosos o fases tempranas del CP [11]. Los NP pueden agruparse también, según su ubicación y contexto, en tres categorías adicionales (Figura 1):

---

[11] Maria D. MARTIN, *et al.* «Lung-RADS: Pushing the limits». En: *Radiographics* 37.7 (oct. de 2017), págs. 1975-1993. DOI: 10.1148/rg.2017170051.

- **Yuxtapleurales:** Pequeños nódulos sólidos localizados cerca de las fisuras pulmonares o en la superficie pleural, los cuales suelen ser benignos y presentan una incidencia aproximada del 21 % [12].
- **Yuxtavasculares:** Nódulos adheridos a los vasos sanguíneos, cuya detección resulta complicada debido a su tonalidad similar a la de los propios vasos. Presentan una incidencia aproximada del 48.7 % y una alta probabilidad de ser malignos [13][14].
- **Aislados:** Nódulos rodeados por tejido pulmonar sin otras anomalías, en su mayoría benignos [15].

Por otra parte, las características morfológicas de los bordes nodulares desempeñan un papel esencial en la clasificación de los NP, ya que facilitan su identificación y permiten diferenciar entre lesiones benignas y malignas con base en patrones visuales identificables en las imágenes (Figura 2). La espiculación es una de las señales más predictivas de malignidad, observándose en más del 90 % de los nódulos malignos y en menos del 10 % de los benignos [16]. La lobulación, asociada a contornos periféricos irregulares, aparece en aproximadamente el 75 % de los nódulos malignos y en apenas 15–25 % de los benignos

---

[12] O. M. METS, *et al.* «Incidental perifissural nodules on routine chest computed tomography: lung cancer or not?» En: *European Radiology* 28 (2018), págs. 1095-1101.

[13] R. HAO; Y. QIANG; X. YAN, *et al.* «Juxta-vascular pulmonary nodule segmentation in PET-CT imaging based on an LBF active contour model with information entropy and joint vector». En: *Computational and Mathematical Methods in Medicine* 2018 (2018).

[14] B. LI, *et al.* «Detection of pulmonary nodules in CT images based on fuzzy integrated active contour model and hybrid parametric mixture model». En: *Computational and Mathematical Methods in Medicine* 2013 (2013).

[15] Patrice MONKAM, *et al.* «Detection and Classification of Pulmonary Nodules Using Convolutional Neural Networks: A Survey». En: *IEEE Access* 7 (2019), págs. 78075-78091. DOI: 10.1109/ACCESS.2019.2920980.

[16] Samuel G ARMATO III, *et al.* «The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans». En: *Medical physics* 38.2 (2011), págs. 915-931.

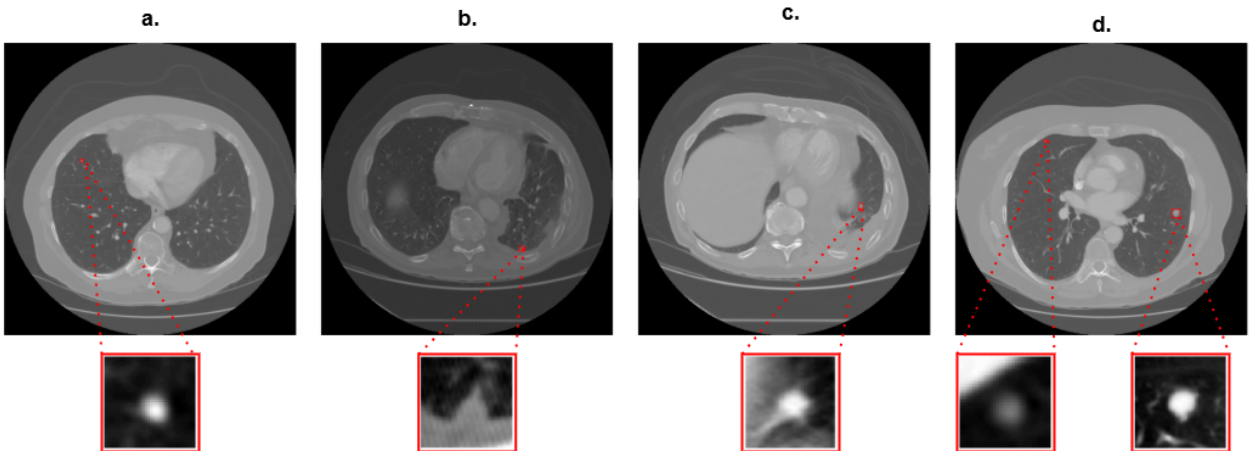

Figura 1. Diferentes tipos de nódulos según su ubicación, contexto espacial y tamaño: a) Nódulo aislado. b) Nódulo yuxtapleural. c) Nódulo yuxtavascular. d) Variabilidad de tamaños entre nódulos pulmonares.

[17]. En contraste, la esfericidad —indicativa de formas ovaladas o regulares— se asocia con benignidad, predominando en cerca del 70 % de los nódulos no cancerosos [18]. Así mismo, la calcificación, especialmente cuando es central, es un fuerte indicador de benignidad, presente en el 50–60 % de estos casos, mientras que su aparición en nódulos malignos es inferior al 10 % [19].

Esta información contextual (las estructuras adyacentes al NP y su descripción morfológica interna) puede tener un papel importante en la detección y caracterización del CP al proporcionar información relevante para discriminar entre nódulos y otras estructuras anatómicas de apariencia similar, como los vasos sanguíneos.

---

[17] Annette MCWILLIAMS, *et al.* «Probability of cancer in pulmonary nodules detected on first screening CT». En: *New England journal of medicine* 369.10 (2013), págs. 910-919.

[18] Sarah J VAN RIEL, *et al.* «Observer variability for classification of pulmonary nodules on low-dose CT images and its effect on nodule management». En: *Radiology* 277.3 (2015), págs. 863-871.

[19] Alexis LACOUT, *et al.* «Pancreatic involvement in hereditary hemorrhagic telangiectasia: assessment with multidetector helical CT». En: *Radiology* 254.2 (2010), págs. 479-484.

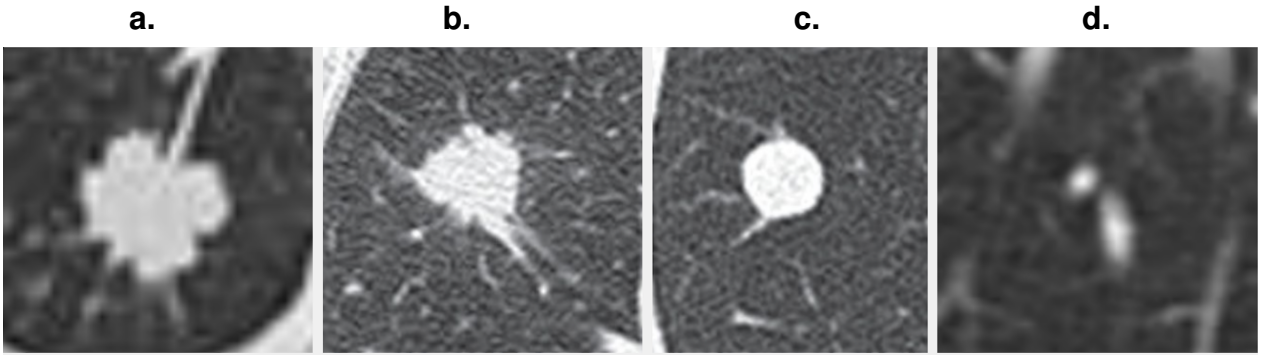

Figura 2. Características morfológicas de los nódulos pulmonares: la lobulación (**a**) y la espiculación (**b**) indican un crecimiento irregular e infiltrante, por lo que se asocian con la presencia de cáncer. Por el contrario, la esfericidad (**c**) y la calcificación (**d**) se relacionan comúnmente con nódulos benignos.

1.2. ESTRATEGIAS DE LOCALIZACIÓN Y REPRESENTACIONES CONTEXTUALES

1.2.1. Modelos basados en atención

Con la irrupción del aprendizaje profundo en la última década, principalmente basados en CNN, el rendimiento de los sistemas de detección de objetos mejoró drásticamente. Sin embargo, las CNNs extraen características a partir de regiones locales de la imagen, lo cual dificulta el aprendizaje de relaciones no locales o de largo alcance entre regiones distantes, pero las cuales pueden estar semánticamente relacionadas. Para superar estas limitaciones, se han propuesto mecanismos de atención que permiten modelar dependencias entre regiones distantes de una imagen, otorgando mayor flexibilidad al modelo para integrar contexto global. El concepto de "atención" fue introducido por primera vez en el campo del procesamiento de lenguaje natural por Bahdanau et al. [20], quienes propusieron un mecanismo para permitir que el modelo enfocara dinámicamente su atención sobre diferentes partes de la secuencia de entrada. Sin embargo, la popularización de esta idea se consolidó con el trabajo de Vaswani et al., quienes introdujeron el modelo *Transformer* y mostraron que, prescindiendo por completo de convoluciones o recurrencias,

[20] Jan K CHOROWSKI, *et al.* «Attention-based models for speech recognition». En: *Advances in neural information processing systems* 28 (2015).

un modelo basado exclusivamente en atención podía alcanzar un rendimiento de estado del arte en traducción automática [21]. Posteriormente, esta arquitectura fue adaptada al dominio visual mediante el Vision Transformer (ViT) propuesto por Dosovitskiy et al. [8]. En los ViT, una imagen es tratada como una secuencia de parches o vectores de características, generalmente obtenidos tras una operación de aplanamiento espacial o mediante un extractor convolucional. Por ejemplo, una imagen de tamaño $H \times W \times C$ puede ser dividida en n parches (o posiciones espaciales) y representada como una secuencia de n vectores en $\mathbb{R}^d$, donde d es la dimensión del espacio embebido. Posteriormente, Carion et al. extendieron este enfoque al problema de detección de objetos con el modelo DETR[22], que emplea un *encoder-decoder* basado en *Transformers* para predecir las ubicaciones y clases de los objetos directamente, sin necesidad de componentes tradicionales como las regiones propuestas o los *anchor boxes* [22]. A continuación se detallan los mecanismos de atención y dos arquitecturas de este tipo usadas para la localización.

Mecanismos de Atención Una vez representada la imagen como una secuencia de parches, es posible aplicar distintos mecanismos de atención sobre esta representación para modelar relaciones contextuales entre los elementos. En el estado del arte, se han desarrollado distintas variantes del mecanismo de atención, entre las que destacan la auto-atención o *self-attention*, la atención cruzada o *cross-attention* y la atención multicabeza o *multi-head attention*. En la auto-atención, se generan representaciones diferentes de la entrada que luego se operan entre sí. Lo anterior causa que cada elemento de la entrada se relacione consigo mismo y con todos los demás elementos de la misma secuencia. Dado un conjunto de vectores de entrada $X \in \mathbb{R}^{n \times d}$, se proyectan tres matrices: vectores

---

[21] Ashish VASWANI, *et al.* «Attention is all you need». En: *Advances in neural information processing systems* 30 (2017).

[22] Nicolas CARION, *et al.* «End-to-end object detection with transformers». En: *European conference on computer vision*. Springer. 2020, págs. 213-229.

de consulta o *queries* $Q = XW_Q$, claves o *keys* $K = XW_K$ y valores o *values* $V = XW_V$, donde $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ son matrices aprendibles. El resultado de la atención se calcula como:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Este mecanismo permite que cada vector de salida sea una combinación ponderada de todos los vectores de entrada, con pesos que reflejan su relevancia relativa. Las matrices Q, K y V provienen de fuentes distintas; por ejemplo, en arquitecturas tipo encoder-decoder, los *queries* suelen derivarse de la representación generada por el decodificador (entrada actual del modelo), mientras que los *keys* y *values* provienen de la representación generada por el codificador (entrada original procesada previamente). Por otra parte, la atención multi-cabeza consiste en replicar varias veces el mecanismo anterior en paralelo, usando diferentes proyecciones lineales, lo cual permite que el modelo aprenda a capturar relaciones desde múltiples subespacios de atención:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

con

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

donde $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_k}$ y $W^O \in \mathbb{R}^{hd_k \times d}$. Este diseño permite al modelo capturar distintos tipos de relaciones entre los elementos de entrada, al operar sobre múltiples subespacios de atención que aprenden patrones de interacción complementarios y heterogéneos.

DETR Los detectores tradicionales de objetos dependen de procesos manuales como la generación de anclas y la supresión de predicciones redundantes (NMS), lo que intro-

duce complejidad y limita el entrenamiento de extremo a extremo [23]. DETR (*DEtection TRansformer*) replantea la detección como una predicción directa de conjuntos, eliminando la propuesta de regiones, anclas y NMS. Para ello, combina una red convolucional (CNN) que extrae un mapa de características, con un *Transformer encoder-decoder*, donde los mecanismos de atención permiten modelar relaciones globales. El *encoder* aplica auto-atención para capturar las interacciones entre todas las regiones, mientras que el *decoder*, usando un conjunto fijo de *object queries*, asocia las predicciones directamente a los objetos, también mediante atención. Igualmente, se presenta la atención cruzada como el mecanismo que une los módulos *transformer encoder* y *transformer decoder*. Además, DETR utiliza una pérdida de asignación bipartita (*Hungarian loss*) que fuerza una correspondencia única entre predicciones y las etiquetas, eliminando duplicados sin necesidad de postprocesamiento. DETR estableció un avance clave hacia una detección de objetos global, simple y de extremo a extremo. No obstante, presenta limitaciones importantes, como su lenta convergencia durante el entrenamiento y su bajo rendimiento en la detección de objetos pequeños, lo que dificulta su adopción en aplicaciones que requieren eficiencia computacional o alta sensibilidad a estructuras sutiles, como ocurre en el análisis de imágenes médicas.

RT-DETR Una evolución reciente de la arquitectura DETR es RT-DETR (Real-Time DETR), la cual aborda de manera directa su principal limitación relacionada con la lenta convergencia y el alto costo computacional durante el entrenamiento [24]. Aunque DETR demostró un rendimiento notable en tareas de detección, su capacidad para localizar objetos pequeños resultó limitada, lo que afecta su aplicabilidad en dominios como el análisis

---

[23] Shifeng ZHANG, *et al.* «Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection». En: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, págs. 9759-9768.

[24] Yian ZHAO, *et al.* «Detrs beat yolos on real-time object detection». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 16965-16974.

de imágenes médicas, donde estructuras sutiles y de reducido tamaño, como los nódulos pulmonares, son clínicamente relevantes. RT-DETR mejora este aspecto al incorporar mecanismos jerárquicos más efectivos para representar detalles locales, manteniendo además una mayor eficiencia durante la inferencia. Utilizando una combinación de atención global y técnicas avanzadas de entrenamiento, logra realizar detecciones en tiempo real sin sacrificar precisión en la localización de objetos pequeños. Además, introduce módulos especializados que optimizan tanto la extracción de características como la selección de consultas, lo que lo convierte en un modelo más ligero y robusto frente a escenarios complejos, con ruido o bajo contraste.

RT-DETR está compuesto por varias etapas clave. En primer lugar, la imagen de entrada es procesada por un backbone convolucional que extrae mapas de características jerárquicos a distintas escalas (S3, S4, S5). Estos mapas son fusionados dentro del módulo *Efficient Hybrid Encoder*, que combina mecanismos como AIFI (*Adaptive Interaction Feature Integration*) y CCFF (*Cross-Scale Context Fusion Fusion*), facilitando la integración de información espacial y contextual entre niveles de resolución. A diferencia de DETR, que inicializa los *object queries* de forma aleatoria, RT-DETR emplea un mecanismo denominado *Uncertainty-minimal Query Selection*, que permite generar consultas embebidas directamente a partir de la salida del encoder. Este enfoque prioriza regiones con alta relevancia semántica y menor incertidumbre, lo que reduce la cantidad de consultas necesarias y acelera la convergencia. En este caso, el decodificador y la cabeza de predicción combinan las consultas seleccionadas con las características de la imagen y las posiciones embebidas para producir predicciones precisas tanto de clase como de ubicación. El modelo también incorpora capas convolucionales eficientes (Conv1$\times$1 y Conv3$\times$3) con activación SiLU y normalización por lotes, lo que mejora la capacidad de refinamiento en la etapa de predicción.

1.2.2. Modelos fundacionales Los modelos de aprendizaje profundo enfrentan desafíos importantes, entre ellos la necesidad de grandes volúmenes de datos para capturar

adecuadamente la variabilidad del problema y garantizar robustez frente al ruido presente en los datos de entrada. Una solución emergente a estos desafíos son los modelos de aprendizaje fundacionales (MF) [25]. Estos modelos están diseñados para adaptarse a una amplia variedad de tareas, ya que se entrenan con grandes cantidades de datos provenientes de múltiples fuentes, y atraviesan distintas etapas de aprendizaje. En su fase inicial, emplean esquemas no supervisados, sin depender de etiquetas de alto nivel, lo que les permite capturar representaciones generales y robustas de bajo nivel. Como resultado, estos modelos muestran un mejor desempeño frente a la variabilidad intra e interclase, y una notable capacidad de generalización y adaptación a distintos dominios y aplicaciones específicas.

Algunos ejemplos de modelos fundacionales son BERT [26], SAM [27], GPT-4 [28], YOLO-World [29] y Grounding DINO [30]. Estos modelos se basan en dos principios fundamentales: el aprendizaje por transferencia (transfer learning, en inglés) y escala del aprendizaje. El aprendizaje por transferencia ha sido un concepto clave desde los inicios del aprendizaje profundo; no obstante, los MF se diferencian de estos por el conjunto de datos de entrenamiento a gran escala [25]. Para alcanzar esta escala, es importante no depender de etiquetas. Por lo cual, estos modelos implementan enfoques no supervisados, proporcionando flexibilidad y eliminando el costo asociado a los datos etiquetados. Además,

---

[25] Rishi BOMMASANI, *et al. On the Opportunities and Risks of Foundation Models*. Ago. de 2021.

[26] Jacob DEVLIN, *et al.* «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *arXiv (Cornell University)* (ene. de 2018). DOI: 10.48550/arxiv.1810.04805.

[27] Alexander KIRILLOV, *et al.* «Segment Anything». En: *arXiv preprint arXiv:2304.02643* (2023).

[28] OPENAI; Steven ADLER; Sandhini AGARWAL, *et al. GPT-4 Technical Report*. Inf. téc. OpenAI, 2024.

[29] Tianheng CHENG, *et al.* «YOLO-World: Real-Time Open-Vocabulary Object Detection». En: *arXiv (Cornell University)* (ene. de 2024). DOI: 10.48550/arxiv.2401.17270.

[30] Shilong LIU, *et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. Mar. de 2023.

gran parte de los modelos fundacionales incluyen módulos específicos que facilitan la interacción con el usuario, ya sea mediante el uso de lenguaje natural o a través de la asignación de áreas de interés utilizando cuadros delimitadores, puntos o máscaras. Este tipo de interacción se conoce como *prompt*. La aplicación de estos modelos puede ofrecer ventajas en varios campos donde los datos específicos para ciertas tareas son escasos. Esto resulta particularmente importante en el área de las imágenes médicas, donde la privacidad de los pacientes y la difícil tarea de establecer acuerdos de cooperación y transferencia de datos son puntos críticos. A continuación se detallan algunos modelos fundacionales que fueron de interés para este proyecto.

Grounding DINO Uno de los principales desafíos en el desarrollo de modelos de detección más flexibles es la capacidad de reconocer e identificar objetos que no pertenecen a un conjunto cerrado de etiquetas previamente definidas. Para abordar este reto, surge el término *grounding*, que se refiere a la asociación explícita entre conceptos expresados en lenguaje natural y regiones específicas dentro de una imagen. Bajo este enfoque, se ha desarrollado *Grounding DINO*, un modelo fundacional de detección de objetos capaz de localizar elementos visuales a partir de expresiones lingüísticas, ya sea mediante nombres de categoría o descripciones referenciales [30]. Su arquitectura se basa en una integración profunda de información visual y textual, y aprovecha un preentrenamiento a gran escala con datos multimodales que vinculan imágenes con descripciones textuales, lo que le permite generalizar eficazmente a nuevos conceptos no vistos durante el entrenamiento. La arquitectura de Grounding DINO combina un codificador multimodal y un decodificador basado en Transformers, donde tanto el texto como la imagen se procesan en paralelo para luego interactuar a través de módulos de atención cruzada. El modelo utiliza un mecanismo de autoatención multimodal reforzado por atención deformable, lo que permite alinear representaciones lingüísticas con regiones específicas de la imagen. Además, incorpora una capa de predicción que realiza regresión de cajas delimitadoras y cálculo de similitud textual de manera conjunta, optimizada mediante una función de pérdida

unificada que combina localización y asociación semántica. Esta estructura le permite no solo detectar objetos conocidos, sino también localizar conceptos descritos libremente en lenguaje natural sin necesidad de reentrenamiento adicional.

Además, la arquitectura utiliza una red *Swin Transformer* para imágenes y *Bidirectional Encoder Representations from Transformers* (BERT) para texto, permitiendo en conjunto procesar y representar tanto imágenes como lenguaje de manera efectiva. En cuanto al Swin Transformer, es una arquitectura jerárquica basada en Transformer que se adapta eficientemente a imágenes mediante ventanas no solapadas de autoatención local. A través de un mecanismo de desplazamiento de ventanas (*window shifting*), permite el modelado de relaciones espaciales globales sin un costo computacional elevado. Es utilizado como *backbone* visual en Grounding DINO para extraer características multiescala de las imágenes. Posteriormente, dichas características se refinan con atención deformable y se fusionan con información textual a través de módulos de atención cruzada (imagen a texto y texto a imagen). Esta estructura permite capturar dependencias espaciales de largo alcance y mejorar la detección de objetos, incluso en escenarios de referencia textual o detección de objetos abiertos.

En cuanto al modelo BERT se basa en una arquitectura de codificadores Transformer profundos con auto-atención (*self-attention*) en múltiples capas [26]. A diferencia de modelos unidireccionales, BERT aprende representaciones bidireccionales del lenguaje, ya que durante su preentrenamiento considera el contexto completo de una palabra en ambas direcciones. Se entrena mediante dos tareas: predicción de palabras enmascaradas y predicción de la siguiente oración, lo que le permite capturar relaciones semánticas complejas. En Grounding DINO, BERT se utiliza como *text backbone* para codificar las entradas textuales que luego se alinean con características visuales a través de mecanismos de atención cruzada. Para mejorar la combinación de ambas modalidades, implementa un módulo que utiliza mecanismos de atención que conectan y enriquecen las representaciones visuales con información textual y viceversa. Además, emplea un método guiado por el

lenguaje que selecciona los elementos visuales más relevantes según el texto de entrada, estableciendo puntos de referencia para el módulo que finalmente realiza predicciones precisas sobre la ubicación y las etiquetas de los objetos.
Grounding DINO se entrena previamente con conjuntos de datos diversos, como COCO (118.000 imágenes con 80 clases), Objects365 v1 y v2 (600.000 y 1.7 millones de imágenes respectivamente con 365 clases), y OpenImages (aproximadamente 9 millones de imágenes con más de 19.000 clases), lo que le permite aprender a relacionar conceptos visuales con descripciones de manera eficiente. Este enfoque ha mostrado resultados destacados en pruebas comparativas como COCO y LVIS, especialmente en la detección sin entrenamiento previo (*zero-shot*) de nuevas categorías, y ha logrado un rendimiento superior en tareas donde los objetos se describen con atributos específicos. A diferencia de otros detectores de conjunto abierto, se destaca por su integración profunda de modalidades en todas las fases del proceso y por el uso de descripciones más precisas y específicas del texto, lo que contribuye a una mejor generalización y un rendimiento constante en diversos escenarios.

YOLO-World La detección de objetos en lenguaje abierto ha impulsado el desarrollo de modelos más adaptables y semánticamente conscientes, entre los cuales destaca *YOLO-World*, una extensión de la familia YOLO diseñada para el reconocimiento de objetos especificados mediante lenguaje natural [29]. YOLO-World permite identificar objetos basados en descripciones textuales o nombres de categorías proporcionadas por el usuario, lo que lo convierte en un detector de vocabulario abierto. El modelo toma como base las arquitecturas YOLOv6 y YOLOv8, conocidas por su eficiencia en tareas de detección en tiempo real, y las expande con mecanismos multimodales que integran lenguaje e imagen. El núcleo de la arquitectura de *YOLO-World* es el módulo *RepVL-PAN (Representative Visual-Language Path Aggregation Network)*, responsable de integrar representaciones visuales y lingüísticas de manera efectiva. Este módulo combina las características extraídas por el *backbone* visual con representaciones semánticas del texto —ya sea en forma

de palabras clave, frases u oraciones completas— permitiendo al sistema alinear regiones específicas de la imagen con conceptos expresados en lenguaje natural, incluso si esas categorías no estuvieron presentes durante el entrenamiento.

La fusión entre modalidades se lleva a cabo mediante dos mecanismos complementarios dentro de RepVL-PAN: el *Text-guided CSPLayer* y el *Image-Pooling Attention*. El primero incorpora información textual directamente en las representaciones visuales de múltiples escalas, utilizando una atención basada en max-sigmoid que permite adaptar las características visuales al contexto lingüístico. Por su parte, el segundo mecanismo refuerza las representaciones textuales a partir de información visual obtenida mediante max pooling, lo que enriquece el entendimiento semántico del texto en función del contenido visual. Esta bidireccionalidad asegura una alineación robusta entre descripciones y regiones relevantes de la imagen.

Durante la inferencia, YOLO-World permite preprocesar y reparametrizar el vocabulario textual como parte de los pesos del modelo. Esto elimina la necesidad de un codificador textual durante la ejecución, lo que reduce significativamente el tiempo de inferencia sin comprometer la precisión.

Para entrenar esta arquitectura, se emplea un esquema de preentrenamiento a gran escala basado en pares región-texto, los cuales asocian directamente áreas específicas de las imágenes con descripciones lingüísticas. Esta estrategia se optimiza mediante una pérdida contrastiva, que maximiza la similitud entre regiones correctas y sus textos correspondientes, mientras penaliza emparejamientos incorrectos. Este objetivo fomenta una mejor discriminación entre clases visualmente similares y mejora la capacidad del modelo para generalizar a nuevas categorías, habilitando el reconocimiento en modo *zero-shot*. Para entrenar el modelo, se emplea un preentrenamiento a gran escala usando pares de imagen-texto, extraídos de datasets diversos y ricos en contenido semántico. A diferencia de enfoques tradicionales, YOLO-World utiliza un objetivo de aprendizaje contrastivo a nivel de región, que fuerza al modelo a asociar correctamente regiones

específicas de la imagen con sus descripciones correspondientes y a distinguirlas de regiones irrelevantes. Esto permite mejorar la discriminación entre conceptos visualmente similares. Gracias a este diseño, YOLO-World puede operar en modo zero-shot, es decir, reconocer clases que no ha visto explícitamente durante el entrenamiento, simplemente a partir de su descripción textual. En conjunto, este enfoque lo posiciona como una solución robusta para tareas complejas de detección donde la interacción con el lenguaje humano es esencial.

1.3. ESQUEMAS COMPUTACIONALES PARA LA LOCALIZACIÓN DE NÓDULOS

Durante los últimos años, se han desarrollado diversos métodos para la detección de nódulos pulmonares en imágenes de tomografía computarizada (TC), con el objetivo de apoyar el diagnóstico temprano del cáncer de pulmón. Entre estos métodos, los enfoques de dos etapas han demostrado ser efectivos al combinar la generación de candidatos y la posterior clasificación para reducir falsos positivos. Por ejemplo, Agnes et al. propusieron un sistema *CADe* de dos etapas que utiliza *Atrous UNet+* para detectar nódulos candidatos en cortes axiales, incorporando *convoluciones dilatadas* y *conexiones de salto* para manejar diversos tamaños. La segunda etapa emplea la red *Pyramid Dilated ConvLSTM (PD-CLSTM)* para clasificar candidatos verdaderos a partir de características espaciales 3D inter- e intra-corte, eliminando falsos positivos [31]. Liao et al. presentaron un método de dos etapas donde la extracción de candidatos se realiza con *PPD-UNet*, una red en forma de U que utiliza *bloques densos* y *pooling paralelo* para una extracción de características robusta. La supresión de falsos positivos se implementa con *BHA-PNet*, una red que toma características profundas de *PPD-UNet* e incorpora *atención híbrida* para mejorar la

[31] S Akila AGNES; J ANITHA y A Arun SOLOMON. «Two-stage lung nodule detection framework using enhanced UNet and convolutional LSTM networks in CT images». En: *Computers in Biology and Medicine* 149 (sep. de 2022), pág. 106059. DOI: 10.1016/j.compbiomed.2022.106059.

percepción espacial y discriminar entre verdaderos y falsos positivos [32]. En otro estudio, Nguyen et al. propusieron *MANet*, un modelo multitarea para detección y segmentación simultánea de nódulos. Este modelo utiliza un *backbone* basado en *UNet* mejorado con *aprendizaje auxiliar de atención multi-rama* para capturar características a diferentes resoluciones. Aunque opera de extremo a extremo, incluye una *generación de candidatos iniciales* y un *paso de refinamiento de propuestas* para reducir falsos positivos [33].

Además, Xu et al. mejoraron un *framework* basado en *Faster R-CNN* para la detección de nódulos, incorporando técnicas como *pirámides de características con aumento de ruta (path augmentation FPN)*, *Online Hard Example Mining (OHEM)*, *función de pérdida GIOU*, *Soft-NMS* y *ROI Align*, así como *convolución deformable (DCN)* [34]. Por otro lado, Manickavasagam et al. propusieron *CNN-5CL*, una red convolucional de 11 capas enfocada en la clasificación de parches de imágenes presegmentados como nódulos o no nódulos, funcionando como la segunda etapa en un sistema *CADe* dedicado a la reducción de falsos positivos tras la generación inicial de candidatos [35].

En cuanto a los modelos de una etapa, Ali et al. presentaron un *framework UNet* eficiente para la detección de nódulos mediante *segmentación semántica de extremo a extremo*, empleando *bloques de convolución dilatada densamente conectados (DCD)* para capturar

---

[32] Miao LIAO, *et al.* «Pulmonary Nodule Detection from 3D CT Image with a Two-Stage Network». En: *International Journal of Intelligent Systems* 2023 (dic. de 2023), págs. 1-14. DOI: 10.1155/2023/3028869.

[33] Tan-Cong NGUYEN, *et al.* «MANet: Multi-branch attention auxiliary learning for lung nodule detection and segmentation». En: *Computer Methods and Programs in Biomedicine* 241 (ago. de 2023), pág. 107748. DOI: 10.1016/j.cmpb.2023.107748.

[34] Jing XU, *et al.* «An improved faster R-CNN algorithm for assisted detection of lung nodules». En: *Computers in Biology and Medicine* 153 (dic. de 2022), pág. 106470. DOI: 10.1016/j.compbiomed.2022.106470.

[35] R. MANICKAVASAGAM; S. SELVAN y Mary SELVAN. «CAD system for lung nodule detection using deep learning with CNN». En: *Medical & Biological Engineering & Computing* 60.1 (nov. de 2021), págs. 221-228. DOI: 10.1007/s11517-021-02462-3.

contexto multiescala [36]. Asimismo, Wu et al. propusieron el modelo *YOLO-MSRF*, basado en *YOLOv7*, que incluye una *Capa de Detección de Objetos Pequeños (SODL)* y un *Módulo de Campo Receptivo Multi-Escala (MSRF)* para mejorar la capacidad de detectar nódulos pequeños y con características borrosas, utilizando también *Convolución Omni-Dimensional Eficiente (EODConv)* [37]. Finalmente, Han et al. propusieron *BiRPN-YOLOvX*, que integra una *red de pirámide de características recursiva bidireccional ponderada (BiRPN)* y una estructura *CBAM CSPDarknet53* para la extracción de características, aplicados dentro de la arquitectura *YOLO* para mejorar la fusión multiescala y la detección directa de nódulos, especialmente pequeños, en una sola etapa [38].

A pesar de los buenos resultados alcanzados por los enfoques basados en una o varias etapas para la detección de nódulos pulmonares, estos métodos tienden a perder información contextual debido a la naturaleza local de las operaciones convolucionales. Para superar esta limitación, se han explorado arquitecturas que integran *mecanismos de atención* y modelos tipo *Transformer*, los cuales permiten capturar relaciones espaciales de largo alcance. Un ejemplo destacado es el trabajo de Mkindu et al., quienes propusieron la arquitectura *3D-NodViT*, basada en *Vision Transformer (ViT)* y *optimización bayesiana*, logrando una detección eficaz de nódulos pulmonares en tomografías computarizadas de tórax con un uso más eficiente de los recursos en comparación con arquitecturas profundas tradicionales [9]. A pesar de los progresos alcanzados con las arquitecturas previamente descritas, la detección precisa de nódulos pulmonares aún enfrenta retos importantes,

---

[36] Zeeshan ALI; Aun IRTAZA y Muazzam MAQSOOD. «An efficient U-Net framework for lung nodule detection using densely connected dilated convolutions». En: *The Journal of Supercomputing* 78.2 (jun. de 2021), págs. 1602-1623. DOI: 10.1007/s11227-021-03845-x.

[37] Xiaosheng WU, *et al.* «YOLO-MSRF for lung nodule detection». En: *Biomedical Signal Processing and Control* 94 (abr. de 2024), pág. 106318. DOI: 10.1016/j.bspc.2024.106318.

[38] Liying HAN, *et al.* «BiRPN-YOLOvX: A weighted bidirectional recursive feature pyramid algorithm for lung nodule detection». En: *Journal of X-Ray Science and Technology* 31.2 (ene. de 2023), págs. 301-317. DOI: 10.3233/xst-221310.

especialmente en lo que respecta a la comprensión del contexto circundante. En este panorama, los *modelos basados en atención* y los *modelos fundacionales* han surgido como enfoques prometedores para abordar estas limitaciones en el análisis de imágenes médicas.

Algunos trabajos recientes han explorado *mecanismos de atención* para mejorar la detección y segmentación de nódulos pulmonares en tomografías computarizadas. El modelo *MCAT-Net* introduce un *módulo de separación de características multiumbral* y un *mecanismo de atención coordinada* para capturar información detallada de bordes y texturas, mejorando la segmentación de nódulos con formas complejas [39]. Asimismo, *TiCNet* combina *módulos de transformadores* con *redes convolucionales 3D* para capturar dependencias a corto y largo alcance, mejorando la detección de nódulos pulmonares en imágenes TC [40]. Por otro lado, *CSEA-Net* incorpora *mecanismos de atención espacial y de canal* para la segmentación precisa de nódulos pulmonares, mostrando un rendimiento sobresaliente en múltiples conjuntos de datos públicos [41]. Además, *MSANet* integra información espacial y de canal para la *detección 3D* de nódulos pulmonares, mejorando la capacidad de extracción de información y la fusión de información multiescala [42].

Recientemente, se ha propuesto el uso de *modelos fundacionales* en imágenes médicas. Por ejemplo, Pai et al. propusieron un *modelo auto-supervisado* para imágenes médicas 3D, el cual fue evaluado en tareas clínicas como *clasificación anatómica*, *predicción de*

---

[39] Tianjiao HU, *et al.* «A lung nodule segmentation model based on the transformer with multiple thresholds and coordinate attention». En: *Scientific Reports* 14.1 (2024), pág. 31743.

[40] Ling MA, *et al.* «TiCNet: transformer in convolutional neural network for pulmonary nodule detection on CT images». En: *Journal of Imaging Informatics in Medicine* 37.1 (2024), págs. 196-208.

[41] Wenhu LIU, *et al.* «CSEA-Net: A channel–spatial enhanced attention network for lung tumor segmentation on CT images». En: *iScience* 28.3 (2025).

[42] Zhitao GUO, *et al.* «Msanet: multiscale aggregation network integrating spatial and channel information for lung nodule detection». En: *IEEE Journal of Biomedical and Health Informatics* 26.6 (2021), págs. 2547-2558.

malignidad y *pronóstico de supervivencia* en cáncer de pulmón. El modelo superó a enfoques supervisados y preentrenados existentes (como *Med3D* y *Models Genesis*), mostrando mayor robustez ante variabilidad y perturbaciones. Además, demostró alta eficiencia y precisión incluso con datos limitados [43]. Si bien estos modelos no han sido ampliamente explorados en la tarea particular de detección de nódulos pulmonares, han mostrado ser capaces de caracterizar correctamente el contexto y la variabilidad de los datos médicos, a pesar de contar con pocas imágenes, por lo que representan una alternativa viable a las redes convencionales de detección.

[43] Suraj PAI, *et al.* «Foundation model for cancer imaging biomarkers». En: *Nature Machine Intelligence* 6.3 (mar. de 2024), págs. 354-367. DOI: 10.1038/s42256-024-00807-9.

2. PROBLEMA DE INVESTIGACIÓN

El cáncer de pulmón es el tipo de cáncer más mortal y prevalente a nivel mundial [1,44]. Las tomografías computarizadas (TC) permiten la localización e identificación de los nódulos pulmonares (NP), que son los principales indicadores del cáncer de pulmón (CP). Sin embargo, la localización de los NP es una tarea fundamentalmente observacional que enfrenta desafíos debido a la alta variabilidad de características morfológicas y contextuales que dificultan su detección [4]. Por ejemplo, los NP varían considerablemente en tamaño, oscilando entre 3 y 30 mm, lo que representa menos del 0.013 % del volumen de las TC. Además, existen similitudes significativas con otras estructuras pulmonares, como los vasos sanguíneos. Estas similitudes y diferencias de tamaño, sumadas a las variadas estructuras torácicas de los pacientes y a la subjetividad de los radiólogos, provocan que hasta un 25 % de los nódulos pasen desapercibidos durante la revisión radiológica [3].

En este escenario, las soluciones computacionales han adquirido un papel central en los sistemas de diagnóstico de CP, aunque aún enfrentan limitaciones importantes. Muchos de estos enfoques, basados principalmente en convoluciones, tienden a perder información contextual visual relevante, lo que disminuye su capacidad para representar adecuadamente la complejidad morfológica y contextual de los NP en las TC. Por otra parte, se han propuesto métodos basados en atención, que permiten capturar relaciones de largo alcance entre las características de la imagen, mejorando así la precisión en la detección de NP. No obstante, estos métodos suelen requerir una mayor cantidad de datos y sus arquitecturas pueden ser susceptibles al sobreajuste. Estas arquitecturas además aún no han explorado exhaustivamente el uso de otra información contextual, que pueda provenir de factores de riesgo o descripciones textuales de los nódulos, lo cual podría aportar en la

---

[44] J. FERLAY, *et al.* «Cancer statistics for the year 2020: An overview». En: *International Journal of Cancer* 149.4 (2021), págs. 778-789.

tarea de detección.

2.1. PREGUNTA DE INVESTIGACIÓN

¿Cómo contribuyen las arquitecturas que incluyen información contextual en la localización de nódulos pulmonares?

3. OBJETIVOS

3.1. OBJETIVO GENERAL

Implementar una representación de atención profunda que aproveche la información contextual para brindar soporte en la tarea de localización de nódulos pulmonares.

3.2. OBJETIVOS ESPECÍFICOS

- Seleccionar un conjunto de datos de secuencias de tomografía computarizada que cuente con etiquetas de localización de nódulos pulmonares.
- Implementar un modelo para la localización de nódulos pulmonares, que explote relaciones espaciales globales mediante mecanismos de atención sobre la imagen.
- Implementar un modelo fundacional adaptado para la localización de nódulos pulmonares, que incorpore tanto contexto visual como de texto.
- Evaluar el rendimiento del método propuesto en cuanto a su capacidad para detectar nódulos pulmonares, usando métricas de localización como *Average Precision* y curva FROC.

4. MÉTODOLOGÍA PROPUESTA

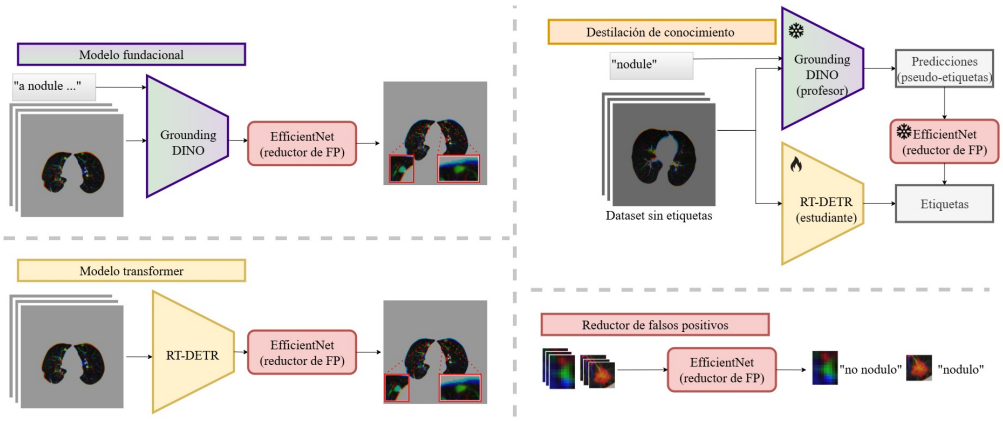

Figura 3. Esquema general de la metodología implementada en este trabajo. Se muestran las dos arquitecturas evaluadas: (i) el modelo *RT-DETR* para la localización de nódulos, y (ii) el modelo fundacional *Grounding DINO*, que integra anotaciones radiológicas como contexto. Adicionalmente, se incluye el módulo reductor de falsos positivos basado en *EfficientNet*, encargado de filtrar predicciones erróneas. Finalmente, se ilustra el esquema de destilación *Profesor–Estudiante*, que transfiere conocimiento del modelo fundacional al modelo compacto para escenarios con recursos limitados.

En este trabajo se exploraron dos arquitecturas basadas en modelos de atención para usar información contextual en el problema de localización de nódulos. Una primera arquitectura adoptó un método transformer (RT-DETR) para localización de nódulos. En una segunda arquitectura, a una escala de parámetros mucho mas grande, se validó el modelo fundacional (Grounding-DINO), explorando relaciones de contexto desde datos de anotaciones radiológicas. Asimismo, se incorporó un módulo reductor de falsos positivos basado en *EfficientNet*, encargado de discriminar entre NP y estructuras anatómicas similares, fortaleciendo la calidad de las predicciones. Por último, se implementó un esquema de destilación desde el modelo fundacional hacia el modelo transformer para abordar el problema en escenarios con capacidades limitadas de infraestructura. La metodología seguida en este trabajo se ilustra en la figura 3 y a continuación se detallan los esquemas implementados.

4.1. MODELO BASADO EN ATENCIÓN PARA LA DETECCIÓN DE NP

Para aprovechar el contexto espacial, en este trabajo, en primera instancia, se implementó un modelo *transformer*. En este caso se implementó y adaptó el modelo RT-DETR [24], para la detección de nódulos pulmonares. Este es un modelo de detección (*end-to-end*) combina detectores tipo YOLO con las ventajas estructurales de los modelos basados en *transformers*. Un esquema de la arquitectura se ilustra en la Figura 4.

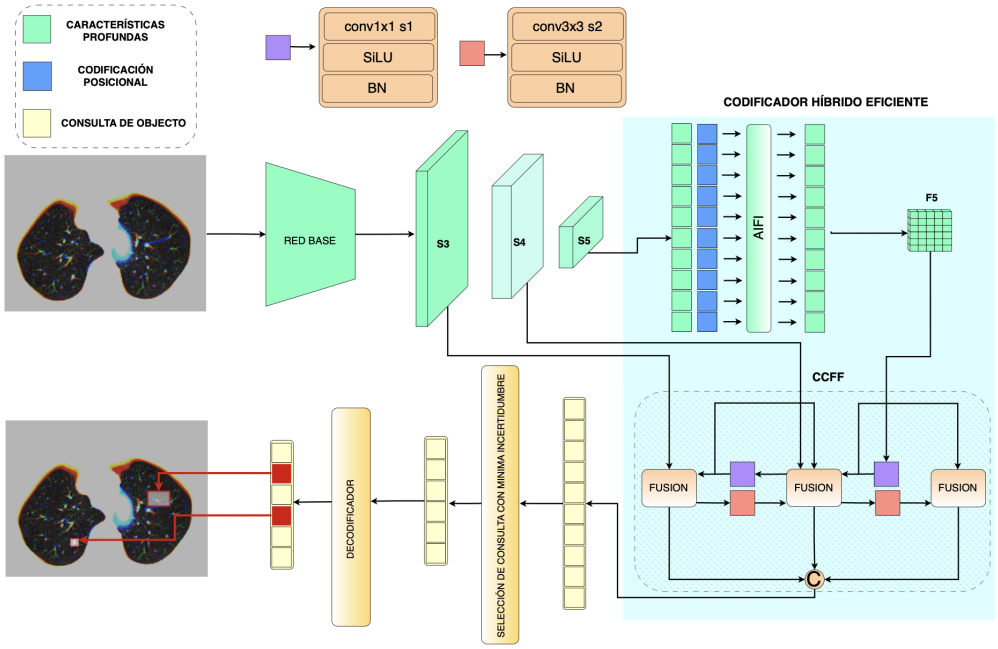

Figura 4. Esquema general de la arquitectura RT-DETR utilizada para la detección de nódulos pulmonares. A partir de características de diferentes escalas, el modelo integra información no local y preserva la información de las diferentes representaciones. Ademas, los *queries* iniciales son extraídos directamente de las características del *encoder*, es decir, son dependientes de la imagen.

Particularmente, cada imagen de entrada $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ se proyecta a través de un banco de filtros de la ResNet-50, representando la entrada como un conjunto de características $F^l = f_l(F^{l-1}),\ l = 1,\ldots,L$, con $F^0 = \mathbf{I}$ y donde L denota el número total de capas de procesamiento. Desde esta proyección, se definen múltiples escalas de representación (características extraídas de diferentes niveles), definidas como $S_i = F^{l_i} \in \mathbb{R}^{H_i\times W_i\times C_i},\quad i \in \{3,4,5\}$, generando tres bloques de características $\{S_3, S_4, S_5\}$,

que representan la información a diferentes escalas (ver Figura 4). Estos bloques multiescala son transformados por un módulo codificador híbrido eficiente (*Hybrid Efficient Encoder*), involucrando la atención intraescala (AIFI) y la fusión cruzada de escalas (CCFF). Luego, la salida de estos módulos es mapeada a un módulo de selección de *queries* con mínima incertidumbre (UMQS, por sus siglas en inglés).

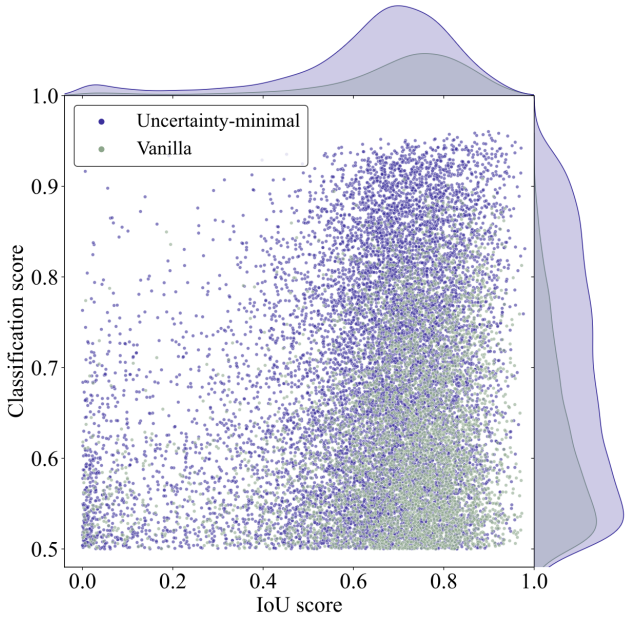

Figura 5. Los puntos verdes representan las características seleccionadas con el esquema tradicional (Vanilla), mientras que los puntos morados corresponden a las características seleccionadas con el esquema propuesto. El análisis revela que los puntos morados se concentran en la parte superior derecha de la gráfica, indicando una mayor calidad de las consultas en términos de clasificación (eje Y) y localización (IoU, eje X).

1. **Atención intra-escala (AIFI: Attention-based Intra-scale Feature Interaction):** La atención intra-escala se aplica únicamente sobre el nivel más alto de características (S_5), ya que este contiene la información semántica más rica. Primero, se aplana y reordena S_5 para ser procesado por un bloque tipo *Transformer*: $Q = K = V =$ Flatten(S_5), $F_5 =$ Reshape(AIFI(Q, K, V)). Donde la atención propia de *Transformer* se define como: Attention$(Q, K, V) =$ softmax $\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$. Aquí, $Q, K, V \in \mathbb{R}^{N \times d}$, con

$N = H_5 \cdot W_5$, y d es la dimensión de los embebidos. En resumen, AIFI corresponde al módulo de atención intra-escala de tipo *Transformer*.

2. **Fusión cruzada de escalas (CCFF: CNN-based Cross-scale Feature Fusion).** Posterior al procesamiento AIFI, se realiza la fusión cruzada jerárquica entre S_3, S_4, F_5 empleando bloques convolucionales. Este proceso puede describirse como: $O =$ CCFF $(\{S_3, S_4, F_5\})$. Cada bloque llamado FUSION dentro de CCFF realiza, para dos *conjuntos de características* de entrada X, Y: $Z_1 =$ Conv1x1$(X + Y), \quad Z_2 =$ Conv1x1$(X + Y),\ U =$ RepBlock$(Z_2),$ Salida $= U + Z_1$. El resultado final, O, es fusión desde múltiples escalas utilizado en la selección de *queries* y el decodificador (*decoder*). Este módulo particularmente realiza la fusión cruzada mediante bloques convolucionales, y O representa la salida final del codificador.

3. **Selección de *queries* con mínima incertidumbre (UMQS)**. Luego de procesada la información en el codificador híbrido, estas características se mapean a un módulo UMQS-*uncertainty-minimal query selection* encargado de seleccionar características que minimizan la incertidumbre conjunta en la predicción de clase y localización. Dado un conjunto de características de salida del codificador $\hat{X} \in \mathbb{R}^D$, para cada $\hat{X}$ se obtiene: $P(\hat{X})$ que es la distribución predicha para localización (caja delimitadora o *bounding box*). Por otra parte, $C(\hat{X})$ es la distribución predicha para clasificación (categoría). Se define la *incertidumbre* como la norma entre las distribuciones de localización y clasificación: $U(\hat{X}) = \|P(\hat{X}) - C(\hat{X})\|$, $\hat{X} \in \mathbb{R}^D$. Para seleccionar las K mejores *queries*, se escogen aquellas con menor incertidumbre $U(\hat{X})$ entre todas las características del codificador. Esto asegura que las *queries* iniciales para el decodificador sean aquellas cuyas predicciones de clase y localización son más consistentes entre sí. En la figura 5 se visualiza la efectividad del módulo. Luego, con el fin de optimizar, se incorpora explícitamente la incertidumbre en la función de pérdida, incentivando que las *queries* seleccionadas tengan baja incertidumbre y, por tanto, sean de mayor calidad para la predicción final.

4. **Decodificador (*decoder*) en RT-DETR.** En la etapa final, el decodificador emplean varias capas *Transformer* para refinar iterativamente un conjunto de consultas (*object queries*) y generar las predicciones de objetos (categoría y caja). En este caso, $Q^0 \in \mathbb{R}^{K \times d}$ es el conjunto de K queries (*object queries*) con dimensión d. Tras L capas, cada query Q_i^L genera predicciones mediante cabezas lineales: la clasificación $\hat{c}_i = \mathrm{Linear}_{\mathsf{cls}}(Q_i^L)$ y la regresión de caja $\hat{b}_i = \mathrm{Linear}_{\mathsf{box}}(Q_i^L)$, donde $\hat{c}_i$ es la distribución sobre las clases y $\hat{b}_i$ las coordenadas normalizadas del **bounding box**. La pérdida total se define como $\mathcal{L}(\hat{X}, \hat{Y}, Y) = \mathcal{L}_{\mathsf{box}}(\hat{b}, b) + \mathcal{L}_{\mathsf{cls}}(U(\hat{X}), \hat{c}, c)$, con $\hat{Y} = \{\hat{c}, \hat{b}\}$ las predicciones y $Y = \{c, b\}$ los valores verdaderos, donde $\mathcal{L}_{\mathsf{box}}$ corresponde a la regresión de caja (p.ej., L1 o GIoU) y $\mathcal{L}_{\mathsf{cls}}$ combina la clasificación con el término de incertidumbre.

4.2. MODELO FUNDACIONAL PARA LA DETECCIÓN DE NP

Como objetivo de este trabajo, en cuanto a la exploración de modelos fundacionales a gran escala, para la localización de nódulos pulmonares, en este trabajo se implementó y adaptó el modelo Grounding DINO [30]. Este modelo ha sido evaluado con éxito en diferentes tareas de localización, siendo entrenado desde múltiples conjuntos de datos, lo que permite la localización de objetos en imágenes naturales, guiada por texto. El modelo Grounding DINO, en su versión original, fue entrenado en total con 14,573,818 objetos, observados en 1,566,783 imágenes naturales. En cuanto a la arquitectura, este modelo está inspirado en módulos que integran múltiples mecanismos de atención (estructura de los *transformers*), permitiendo así conservar y codificar patrones complejos, capturando relaciones no locales, contextuales y además aprovechando el carácter multimodal al soportar vectores codificados desde información textual. Así, esta arquitectura está compuesta por diversos módulos diseñados para procesar y relacionar la información visual y el contexto aportado por la información textual para la localización de objetos. Este estudio incorporó diversas entradas textuales, dado que el objetivo central es modelar y validar

la información extraída principalmente de las tomografías, con el fin de localizar nódulos pulmonares (NP). En particular, la Figura 6 muestra un esquema general de la arquitectura Grounding-DINO adaptada en este trabajo y a continuación se describe en detalle cada uno de sus componentes.

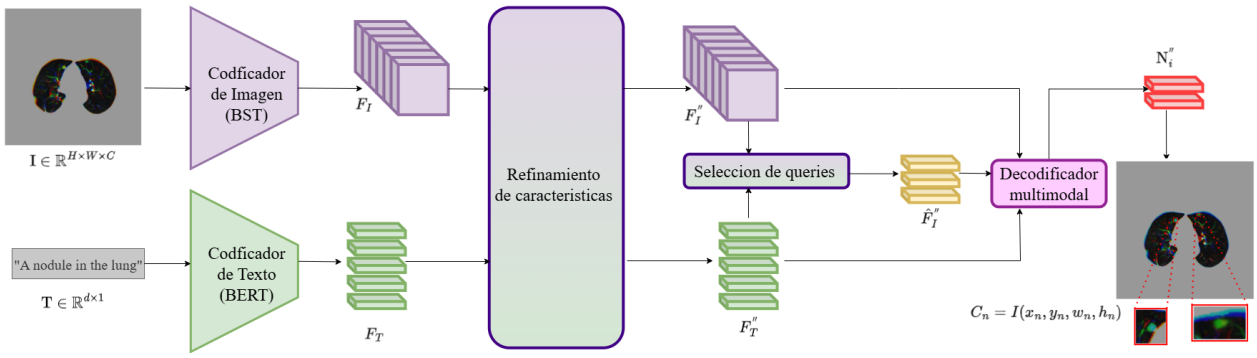

Figura 6. Esquema general de la arquitectura Grounding DINO utilizada para la detección de nódulos pulmonares. La imagen y el texto se codifican por separado para extraer sus características. Luego, se alinean mediante mecanismos de atención cruzada que combinan información visual y textual. A partir de esta fusión, el decodificador genera predicciones de cajas delimitadoras y etiquetas referenciales, permitiendo la detección de objetos guiada por lenguaje natural

Como se ilustra en la figura 6, en la primera etapa, en la rama de procesamiento visual, primero se define un codificador multi-escala no local, que recibe una imagen $I \in \mathbb{R}^{H \times W \times C}$, donde $H \times W$ representan las dimensiones espaciales y, $C = 3$. La entrada I es dividida en n parches, calculados mediante la operación $n = \frac{H}{4} \times \frac{W}{4}$, donde 4 corresponde a las dimensiones de los parches 4×4 utilizados en este trabajo. De esta manera, se obtiene un total de $16{,}384$ parches. Cada uno de estos parches es proyectado en una capa de proyección lineal para obtener un vector embebido por parche (*token*), denotado como $\mathbf{p} = \{p_1, p_2, \ldots, p_n\}$. Esta representación de *tokens* es mapeada a una representación de pares de bloques consecutivos del *Swin Transformer* (BST), expresando cada par como $\beta^i_{\langle 1,2 \rangle}$, siendo i la etapa de procesamiento. Entonces, realizando un procesamiento secuencial, se obtiene una nueva representación de vectores embebidos:

$$F_I = \beta_L^{<1,2>} \circ (\cdots \beta_i^{<1,2>} \circ (\cdots \beta_1^{<1,2>}(I)))$$

En esta representación compuesta F_I, se obtiene un conjunto de vectores embebidos que explotan la información visual, aprovechando el contexto extraído por múltiples módulos Swin Transformer $\beta_i^{<1,2>}$ [45]. Este bloque BST, basado en los módulos de atención Swin Transformer, se detalla en la figura 7. Como se ilustra en la figura, particularmente, el bloque $\beta_i^{<1>}$ emplea el mecanismo de auto-atención sobre ventanas no superpuestas de tamaño fijo, conocido como **W-MSA** (*Window-based Multi-head Self-Attention*), mientras que el bloque $\beta_i^{<2>}$ introduce una mejora clave mediante el mecanismo de **SW-MSA** (*Shifted Window-based Multi-head Self-Attention*). Es así como, el bloque BST utiliza una representación de auto-atención, definida como:

$$MSA = 4hwC^2 + \begin{cases} a = 2(hw)^2C & \text{para el bloque } \beta_i^{<1>} \text{ (W-MSA)} \\ b = 2 \times 7^2 hwC & \text{para el bloque } \beta_i^{<2>} \text{ (SW-MSA)} \end{cases}$$

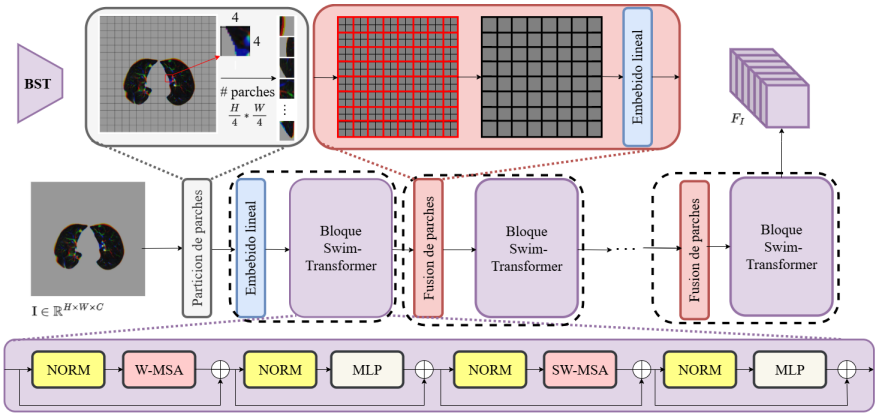

Figura 7. Esquema del codificador de imagen. Este esquema ilustra el proceso de extracción jerárquica de características visuales a través del backbone, el cual emplea bloques Swin Transformer con mecanismos W-MSA y SW-MSA, permitiendo capturar relaciones espaciales locales y globales de manera eficiente.

En este diseño, el uso de ventanas desplazadas en el segundo bloque permite que las

---

[45] Ze LIU, *et al.* «Swin Transformer: Hierarchical Vision Transformer using Shifted Windows». En: *arXiv (Cornell University)* (ene. de 2021). DOI: 10.48550/arxiv.2103.14030.

regiones que inicialmente se encontraban en ventanas separadas ahora se solapen parcialmente en la siguiente etapa. Este desplazamiento facilita la interacción entre parches vecinos que antes no compartían contexto, mejorando la capacidad del modelo para capturar relaciones espaciales más amplias sin necesidad de atención global. Este mecanismo se complementa con un proceso de fusión parcial entre parches, donde a cada vector embebido se le concatena información parcial del vector adyacente, fortaleciendo así la representación regional en cada etapa secuencial de procesamiento.

En cuanto a la rama de procesamiento textual, este modelo emplea un codificador de texto basado en la arquitectura BERT [26] - *Bidirectional Encoder Representations from Transformers*- usando como entrada dos diferentes sentencias: '*A nodule in the lung*' y la palabra '*nodule*'. En la Figura 8 se detalla el funcionamiento del codificador BERT, en donde la sentencia de entrada se representa como una tupla de *tokens* $\varsigma = [Vt \cup Vp \cup Vs]$, que brinda información codificada relacionada con el token t, la posición relativa del token p y la codificación relativa al segmento de la frase, donde cada palabra s está posicionada, respectivamente. Esta representación embebida, con la información contextual tanto del token como de la posición relativa de la frase y su respectivo segmento, es procesada a través de módulos de *auto-atención (Self-Attention (SA))*, obteniendo una representación:

$$F_T = \tau_L^{<1,2>}(\cdots \tau_i^{<1,2>}(\cdots \tau_1^{<1,2>}(\varsigma)))$$

La arquitectura BERT utilizada en esta rama de procesamiento textual fue preentrenada mediante tareas no supervisadas, específicamente el modelado de lenguaje enmascarado (MLM) y la predicción de la siguiente oración (NSP). En la tarea de MLM, ciertos tokens de la secuencia de entrada se enmascaran aleatoriamente, y el modelo debe predecir dichos tokens a partir del contexto que los rodea, lo que favorece una comprensión bidireccional del lenguaje. Por otro lado, la tarea de NSP consiste en presentar al modelo pares de oraciones y requerir que determine si la segunda oración sigue secuencialmente a la primera, fortaleciendo así su capacidad para captar relaciones entre oraciones.

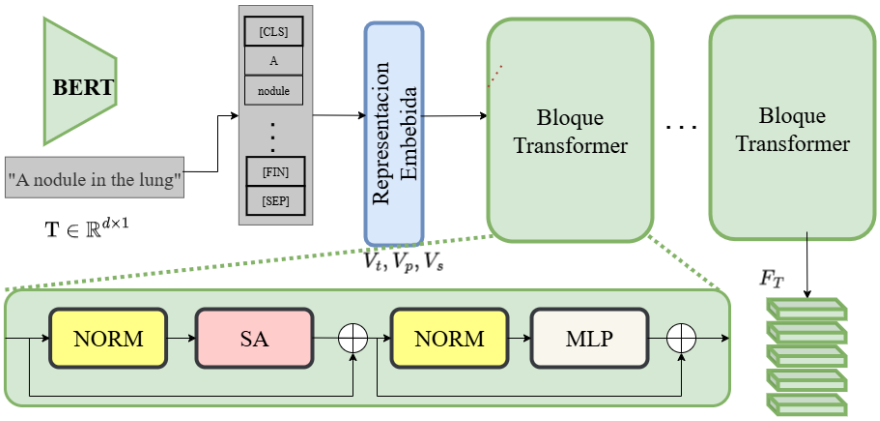

Figura 8. Esquema del codificador de texto. Transforma la sentencia de entrada, en una representación contextual embebida que combina codificación de token, posición y segmento. Estas representaciones son procesadas mediante múltiples capas de autoatención (SA) para capturar relaciones semánticas complejas.

Posteriormente, a partir de las representaciones visuales F_I y textuales F_T, se obtiene un conjunto de vectores embebidos procesados a través de un bloque de refinamiento de características. Un análisis detallado de este bloque de refinamiento se ilustra en la figura 9. Particularmente, este bloque utiliza módulos de atención, obteniendo así un nuevo conjunto de representaciones de imagen F_I'' como de texto F_T''. En particular, este bloque de refinamiento hace uso de módulos de autoatención (SA), atención cruzada (CA) y atención deformable [46]. Inicialmente, las representaciones textuales F_T son mapeadas en un módulo de autoatención $F_T' = SA(Q_{F_T}, K_{F_T}, V_{F_T}) = softmax\left(\frac{Q_{F_T} K_{F_T}}{\sqrt{d_k}}\right) V_{F_T}$, obteniendo vectores F_T' que capturan relaciones en la misma representación (como se ilustra en la Figura 10-b).

Por otra parte, los vectores asociados a la representación visual F_I se integran en un esquema de atención deformable (DA), el cual permite modelar relaciones visuales mediante mecanismos de auto-atención ajustados dinámicamente. Esta adaptación se logra mediante el uso de campos vectoriales denominados *offsets*, que modifican las posiciones

[46] Zhuofan XIA, *et al.* «Vision Transformer with Deformable Attention». En: *arXiv (Cornell University)* (ene. de 2022). DOI: 10.48550/arxiv.2201.00520.

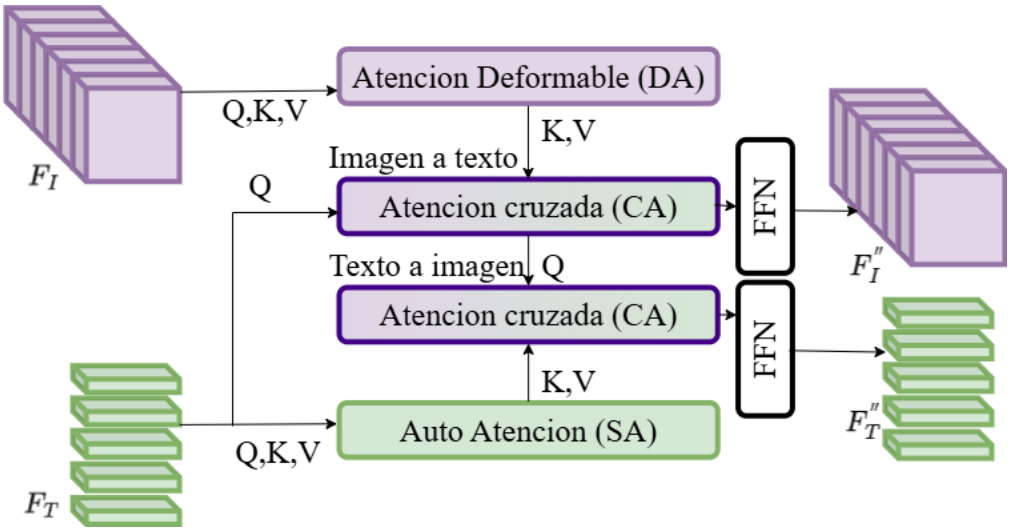

Figura 9. Esquema de módulo de refinamiento de características. A partir de las representaciones iniciales visuales y textuales, se genera un conjunto de vectores embebidos que son refinados haciendo uso de módulos de auto-atención, atención deformable y atención cruzada, obteniendo nuevas representaciones enriquecidas.

estándar de atención para enfocarse en regiones más relevantes de la imagen, mostrando en la literatura una mayor capacidad para capturar estructuras espaciales. El conjunto de estos modelos de atención se presenta en detalle en la Figura 10-a. Particularmente, en este módulo se aprenden vectores de desplazamiento D_P a partir de una representación θ_{offset},que corrige la localización de las representaciones visuales, causada por el proceso convolucional. Entonces, las características de entrada son proyectadas convolucionalmente en el esquema *offset* ($Q_{F_I} = F_I \cdot W_Q$) y son mapeadas en la red $\theta_{offset}(Q_{F_I})$, la cual actualiza las características $\hat{F}$, como $\hat{F} \leftarrow D_P(\Delta P)$, donde $D_P = \theta_{offset}(Q_{F_I}) + R_P$. A partir de esta representación ajustada mediante offsets, se lleva a cabo el proceso de multi-atención, en el cual se aplican proyecciones complementarias de *queries*, *keys* y *values* para capturar relaciones contextuales entre los elementos visuales, realizando proyecciones complementarias $\hat{K}_F = \hat{F} \cdot W_K$ y $\hat{V}_F = \hat{F} \cdot W_V$. Cada módulo de atención individual está definido como: $Z_m = \sigma\left(\frac{Q^m_{F_I} \cdot K^m_{\hat{F}}}{\sqrt{d}}\right) \cdot V^m_{\hat{F}}$, y la representación conjunta de múltiples cabezas que se obtiene mediante la concatenación de las salidas de cada cabeza: $DA = concat(Z_1, Z_2, \ldots, Z_m)$,. Este mecanismo ha demostrado una mayor efectividad para

representar objetos de tamaño reducido, ya que permite capturar de manera más precisa las dependencias contextuales internas dentro de una imagen, mejorando así la sensibilidad del modelo ante estructuras locales complejas, el rendimiento y la generalización en tareas de detección, frente a otros esquemas de atención.

Posteriormente, en el bloque de refinamiento de características se aplican bloques de atención cruzada haciendo uso de las características obtenidas de los módulos SA y DA respectivamente. Estos bloques de atención cruzada se procesan de manera consecutiva como:

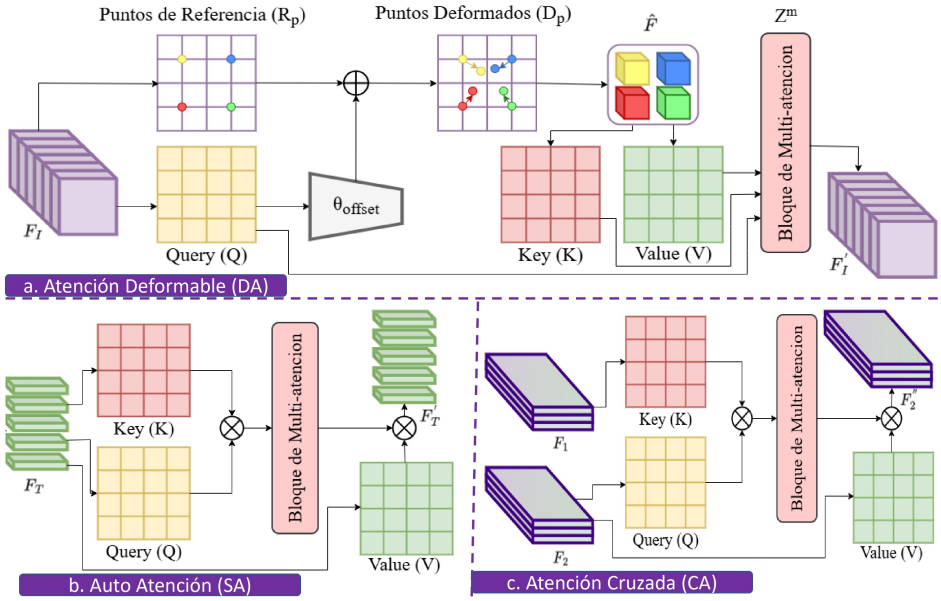

Figura 10. Módulos de atención deformable, auto-atención y atención cruzada: a) Atención deformable que aprende *offsets* espaciales para ajustar la localización visual, b) Auto-atención que modela dependencias internas en la representación textual, c) Atención cruzada que fusiona eficientemente información visual y textual.

$$F_I'' = CA(Q_{F_I'}, K_{F_T}, V_{F_T}) = softmax\left(\frac{Q_{F_I'} K_{F_T}}{\sqrt{d_k}}\right) V_{F_T}$$

donde F_I' representa la característica visual refinada mediante el mecanismo de atención deformable, mientras que F_I'' corresponde a la representación visual enriquecida con

información textual a través de atención cruzada. Análogamente, para la representación textual:

$$F_T^{''} = CA(Q_{F_T^{'}}, K_{F_I}, V_{F_I}) = softmax\left(\frac{Q_{F_T^{'}} K_{F_I}}{\sqrt{d_k}}\right) V_{F_I}$$

donde $F_I^{'}$ representa la característica visual refinada mediante el mecanismo de atención deformable, mientras que $F_I^{''}$ corresponde a la representación visual enriquecida con información textual a través de atención cruzada. Adicionalmente, cada una de estas representaciones pasa por una capa de proyección lineal *(FFN)*.

Seguidamente, estas características refinadas pasan por un módulo de selección de *queries* guiados por lenguaje, cuyo objetivo es seleccionar las representaciones visuales $F_I^{''}$ que mejor coincidan con las representaciones textuales $F_T^{''}$ a través de una comparación de máximos. Este proceso permite identificar correspondencias de alta similitud entre ambas modalidades, utilizando una selección de características con operaciones basadas en máximos que favorecen a las correspondencias alineadas en ambas modalidades. Formalmente, se seleccionan 900 *queries* por defecto de la siguiente manera:

$$\hat{F}_I^{''} = Top_{900}(Max^{(-1)}(F_I^{''} F_T^{''\top}))$$

En esta fase, el conjunto de tres vectores embebidos resultantes $F_I^{''}$, $\hat{F}_I^{''}, F_T^{''}$ se integra en una etapa de decodificación multimodal, cuya finalidad es predecir las coordenadas de las cajas delimitadoras (*bounding boxes*) asociadas a los objetos detectados (Fig. 11).

Este decodificador comienza aplicando un mecanismo de autoatención sobre los vectores provenientes del conjunto de *queries* seleccionados, representados por: $\hat{F}_I^{''} = SA(Q_{\hat{F}_I^{''}}, K_{\hat{F}_I^{''}}, V_{\hat{F}_I^{''}})$. Posteriormente, estos vectores se integran con la representación visual enriquecida $F_I^{''}$ a través de un módulo de atención cruzada, que permite combinar de manera efectiva la redundancia informativa entre ambas representaciones: $N_i = CA(Q_{\hat{F}_I^{''}}, K_{F_I^{''}}, V_{F_I^{''}})$. A continuación, los vectores N_i se enriquecen adicionalmente con información tex-

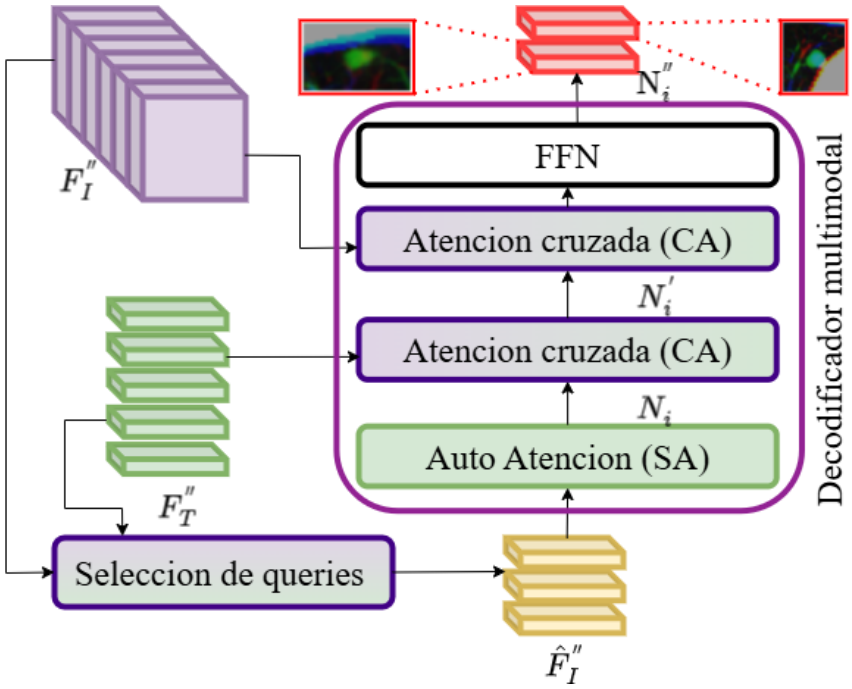

Figura 11. Esquema de la selección de *queries* y decodificador multimodal. Se extraen características visuales con mayor correspondencia a las características textual, que se procesan primero por una capa de auto-atención, seguido de atención cruzada con características visuales y luego con características textuales, y finalmente una capa *FFN* genera los vectores que definen las cajas delimitadoras.

tual mediante otro bloque de atención cruzada, incorporando contexto semántico desde la modalidad textual: $N_i^{'} = CA(Q_{N_i}, K_{F_T^{''}}, V_{F_T^{''}})$. Los *queries* actualizados $N_i^{'}$, que ahora contienen características tanto visuales como textuales, se emplean para predecir las coordenadas de las cajas delimitadoras asociadas a los objetos detectados. Finalmente, una capa *FFN* transforma estas representaciones en un nuevo conjunto de vectores $N_i^{''}$, los cuales codifican directamente los parámetros espaciales de las cajas propuestas: $C_n = I(x_n, y_n, w_n, h_n)$, donde x_n e y_n corresponden a las coordenadas del centro de la caja, mientras que w_n y h_n indican su ancho y altura, respectivamente. Esta salida constituye la predicción final del modelo, generando candidatos a nódulos pulmonares alineados con las características semánticas extraídas de ambas modalidades.

4.3. DESTILACIÓN DE CONOCIMIENTO: ESTRATEGIA PROFESOR-ESTUDIANTE

Como extensión de la investigación desarrollada, en este trabajo se exploró una adaptación que integrara las dos estrategias propuestas. En este sentido, se busca explorar cómo se puede transferir conocimiento desde el modelo fundacional hasta la arquitectura *transformer*, la cual es de menor escala y número de parámetros y podría adaptarse en entornos clínicos más desafiantes. Así, la destilación desde el modelo fundacional podría ser clave para operar en entornos clínicos con recursos limitados.

Particularmente, la destilación de conocimiento (*Knowledge Distillation*, KD), en un esquema Profesor–Estudiante, es una técnica de aprendizaje profundo en la que un modelo grande con mayor valor de generalización transfiere a un modelo más pequeño el conocimiento adquirido, con el fin de obtener modelos compactos que mantienen un rendimiento cercano al del original y que resultan más viables en escenarios con recursos limitados o aplicaciones en tiempo real [47]. En el estado del arte, la destilación ha permitido comprimir

---

[47] Zhaohui ZHENG, *et al.* «Localization distillation for dense object detection». En: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, págs. 9407-9416.

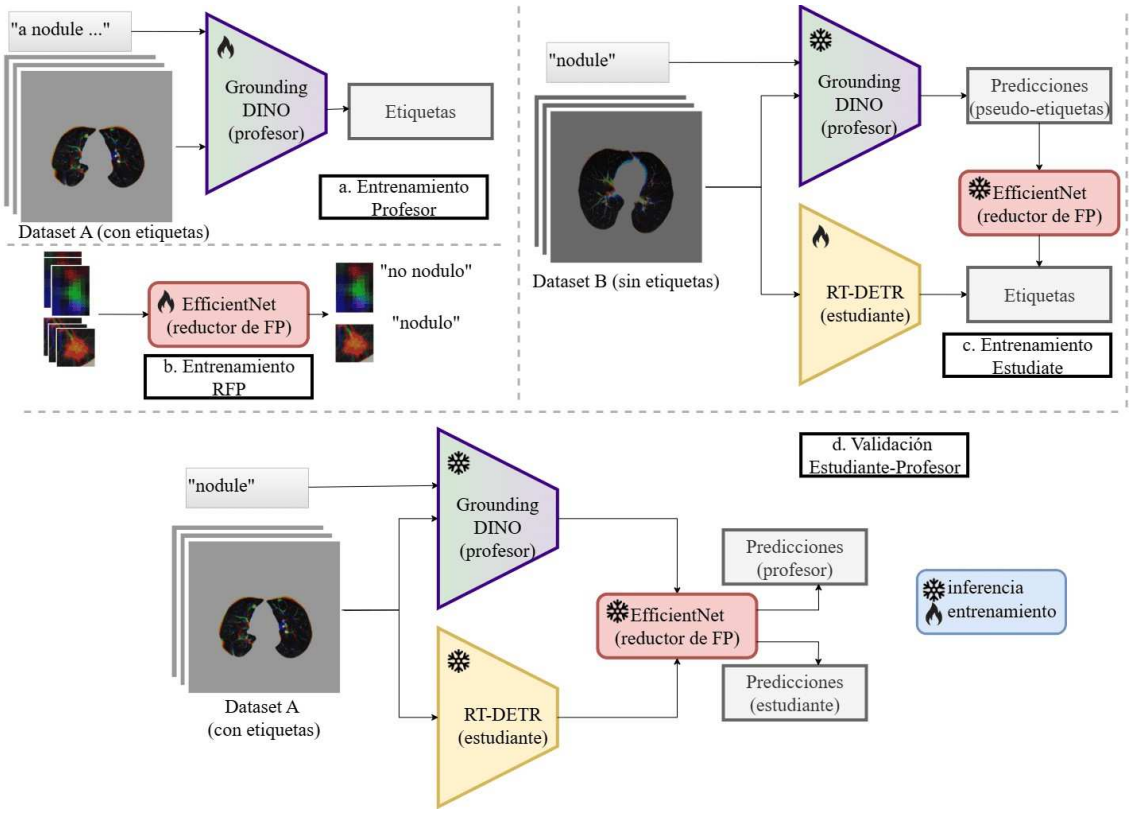

Figura 12. Esquema del método Profesor–Estudiante propuesto. El modelo fundacional (Grounding DINO), entrenado con un conjunto de datos anotado, actúa como profesor generando pseudo-etiquetas sobre un conjunto masivo de imágenes no anotadas. Estas predicciones son filtradas por un reductor de falsos positivos (*EfficientNet*) para conservar únicamente las detecciones confiables, que luego sirven como insumo para entrenar al modelo estudiante (RT-DETR), más ligero y eficiente.

detectores sin grandes pérdidas de precisión, mejorar la detección de objetos pequeños, facilitar el aprendizaje incremental y optimizar el rendimiento en contextos semi-supervisados y multimodales [48]. En contextos de imágenes naturales, algunos trabajos muestran que técnicas específicas como la *Localization Distillation* incrementan el *average precision* del estudiante [47], mientras que enfoques como la imitación del *ranking* de detecciones del

[48] Zhihui LI, *et al.* «When object Detection meets knowledge Distillation: A survey». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.8 (mar. de 2023), págs. 10555-10579.

maestro permiten transferir de forma indirecta el efecto de la supresión de no-máximos [49]. Asimismo, se ha demostrado que en escenarios binarios la destilación contribuye a reducir falsos positivos y calibrar mejor las puntuaciones de confianza [50].

Siguiendo esta línea, en este trabajo se adoptó una variante del paradigma Profesor–Estudiante adaptada al problema de detección de nódulos pulmonares (veáse Figura 12). Se optó por un esquema de pseudo-etiquetado a gran escala: un modelo fundacional de gran capacidad (Grounding DINO) generó predicciones sobre un conjunto masivo de imágenes no anotadas. Estas etiquetas fueron refinadas con un reductor de falsos positivos. Este paso intermedio permitió depurar las detecciones iniciales, conservando únicamente aquellas con mayor grado de confianza, lo que no solo mejoró la calidad de las pseudo-etiquetas, sino que también redujo el ruido en el conjunto de entrenamiento utilizado posteriormente por el modelo estudiante. De este modo, únicamente las detecciones confiables fueron empleadas como insumo para entrenar un modelo más pequeño y veloz basado en mecanismos de atención (RT-DETR). Esta estrategia mantiene la esencia de la destilación al transferir el conocimiento del maestro hacia un estudiante más eficiente, pero se diferencia al incorporar un filtro intermedio que mitiga el ruido de las pseudo-etiquetas, garantizando que el modelo final aprenda de ejemplos de alta calidad y alcanzando un equilibrio entre precisión y eficiencia, condición indispensable en aplicaciones clínicas.

4.4. MODELO REDUCTOR DE FALSOS POSITIVOS - FP -

Las estrategias computacionales para la detección de NP en imágenes de TC enfrentan desafíos en la abundancia de falsos positivos (FP) en sus predicciones. Este problema se

---

[49] Gang LI, *et al.* «Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation». En: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 2. 2022, págs. 1306-1313.

[50] Martin AUBARD, *et al.* «Knowledge distillation in YOLOX-ViT for side-scan sonar object detection». En: *arXiv preprint arXiv:2403.09313* (2024).

debe a la similitud visual de los nódulos con otras estructuras anatómicas, como vasos sanguíneos o ramificaciones bronquiales, así como con artefactos generados por el movimiento del paciente o por defectos en los detectores. Para abordar dichas limitaciones, en este trabajo se implementó un enfoque basado en un modelo convolucional preentrenado llamado *EfficientNet* [51], que permite identificar características propias de los NP, descartando así FPs. Este esquema de reducción de FP fue aplicado a las tres aproximaciones presentadas en este trabajo: el modelo fundacional, la arquitectura *transformer* y la metodología Profesor-Estudiante, que integra las dos anteriores.

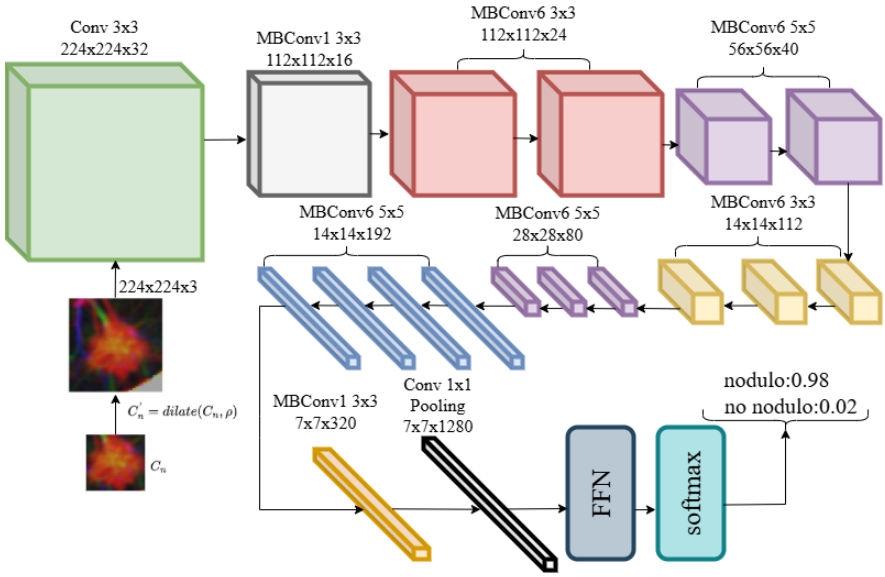

Figura 13. Esquema general de la arquitectura *EfficientNet* implementada para la reducción de falsos positivos. El modelo está conformado por bloques MBConv, que aplican convoluciones invertidas y operaciones de compresión-expansión de canales, permitiendo una extracción eficiente de características discriminativas en los candidatos a NP.

En la Figura 13 se presenta el esquema general de la arquitectura empleada para la reducción de falsos positivos (FPs). En una primera etapa, cada candidato C_n, generado por alguna de las estrategias implementadas en el trabajo, es sometido a un proceso de

[51] Mingxing TAN y Quoc V. LE. «EfficientNet: Rethinking model scaling for convolutional neural networks». En: *arXiv (Cornell University)* (ene. de 2019). DOI: 10.48550/arxiv.1905.11946.

contextualización local. Para ello, las coordenadas de la caja se expanden mediante un factor $\rho \leq 1$, proporcional al tamaño original, lo que se expresa como:$C_n^{'} = dilate(C_n, \rho) = I(x_n, y_n, w_n(1+\rho), h_n(1+\rho))$. Con esta dilatación se incorpora el contexto anatómico circundante al nódulo, favoreciendo su distinción frente a estructuras similares.

Posteriormente, los parches extraídos se normalizan a una dimensión fija $[H^{'}, W^{'}, C]$, donde $H^{'}$ y $W^{'}$ representan las dimensiones espaciales y C corresponde al número de canales, garantizando uniformidad en la entrada de las etapas posteriores del procesamiento. En este trabajo se fijó $H^{'} = 224 \times W^{'} = 224$, dado que el enfoque propuesto requiere mantener una resolución estándar que permita conservar la información contextual del entorno anatómico y, al mismo tiempo, asegurar compatibilidad con arquitecturas previamente entrenadas en este tamaño de entrada. A continuación, se realiza la extracción de características a través de la red *EfficientNet*, que está constituida de bloques MBConv (*Mobile Inverted Bottleneck Convolution*), que actúan de manera independiente sobre cada canal de entrada, seguido de una convolución que combina la información de todos los canales. Adicionalmente, cada bloque integra un módulo de *Squeeze-and-Excitation* (SE), encargado de re-calibrar dinámicamente la importancia de los canales: primero resume la información global de cada canal (*squeeze*) y, posteriormente, asigna pesos de atención que refuerzan las características más relevantes y atenúan las menos útiles (*excitation*). Esto permitió disminuir la complejidad computacional sin afectar la capacidad del modelo para extraer características discriminativas. La arquitectura *EfficientNet* se basa en un esquema de escalado compuesto, que ajusta de manera equilibrada parámetros como la profundidad (número de capas) y el ancho (cantidad de filtros o canales), lo que facilita generar variantes de distinto tamaño, desde *EfficientNet-B0* hasta *B7*, adaptándose a diferentes escenarios de capacidad y recursos. En este trabajo se empleó la versión más compacta, *EfficientNet-B0*, cuya arquitectura ligera resulta especialmente adecuada para conjuntos de datos limitados, como las imágenes médicas de NP, reduciendo el riesgo de sobre-ajuste y favoreciendo un entrenamiento más eficiente en condiciones de escasez de

datos. Así, en la capa fina, mapeando la información a una capa *softmax* se determina la probabilidad de correspondencia a NP o FP.

5. DISEÑO EXPERIMENTAL

5.1. CONJUNTOS DE DATOS

Para el entrenamiento, ajuste y validación de los enfoques propuestos, en este trabajo se consideraron dos conjuntos de datos, los cuales se resumen a continuación:

Conjunto de datos público LIDC-IDRI *(Lung Image Database Consortium and Image Database Resource Initiative)*. Este conjunto de datos contiene 1.018 estudios que incluyen un total de 2.742 nódulos. Además de las imágenes, este conjunto incorpora información y anotaciones relevantes sobre la ubicación, las características de malignidad y la caracterización de los nódulos pulmonares (NP). Estos datos son relevantes para el modelo fundacional. En particular, para ajustar el enfoque fundacional se usaron características como la calcificación, lobulación, espiculación y sutileza, cuyos índices presentan anotaciones numéricas; por ello, se codificó dicha información a formato textual para incorporarla en el modelo [52]. Estas anotaciones fueron realizadas de forma independiente por hasta cuatro radiólogos expertos [16]. Entre las características definidas para los NP se incluyen su tamaño y el nivel de dificultad para su detección según su densidad. La Tabla 1 presenta la distribución del conjunto de datos en los subconjuntos de entrenamiento, validación y prueba.

Conjunto de datos *National Lung Screening Trial* (NLST) [53]. Este conjunto de datos proviene de un estudio clínico a gran escala que incluyó aproximadamente 150,000

---

[52] Pia OPULENCIA, *et al.* «Mapping LIDC, RadLex™, and Lung Nodule Image Features». En: *Journal of Digital Imaging* 24.2 (mar. de 2010), págs. 256-270. DOI: 10.1007/s10278-010-9285-6.

[53] The National Lung Screening Trial Research TEAM. «Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening». En: *New England Journal of Medicine* 365.5 (2011), págs. 395-409. DOI: 10.1056/NEJMoa1102873.

Tabla 1. Distribución del conjunto de datos LIDC-IDRI.

Subconjunto	Porcentaje	Scans	Imagenes	NP
Entrenamiento	70 %	582	1,415	1534
Validación	10 %	85	203	225
Prueba	20 %	162	386	493

exploraciones. Sin embargo, este conjunto de datos presenta una disponibilidad muy limitada de exploraciones con anotaciones de localización de nódulos realizadas por radiólogos, ya que solo 88 de los 2086 *scans* contaban con dichas marcas. Por esta razón, este subconjunto anotado se utilizó exclusivamente en la etapa de validación de los modelos de detección, con el fin de evaluar su capacidad de adaptación a un conjunto de datos distinto.

Para el enfoque Profesor–Estudiante, en el cual el modelo fundacional procesó dicho conjunto de datos para generar pseudo-etiquetas, que luego fueron utilizadas en el entrenamiento del modelo compacto. Posteriormente, los parches obtenidos fueron procesados por el modelo reductor de falsos positivos, lo que permitió depurar las predicciones y fortalecer el conjunto de entrenamiento. Los estudios preprocesados fueron divididos en dos subconjuntos: un 80 % para entrenamiento (1542 *scans* con 11,785 imágenes) y un 20 % para validación (386 *scans* con 2738 imágenes).

Para preparar adecuadamente los datos radiológicos antes de su ingreso al modelo, se realizó la extracción de los cortes (*slices*) a partir de cada volumen de estudio. Posteriormente, las imágenes fueron convertidas al formato PNG y sometidas a un proceso de preprocesamiento basado en un enfoque de contextualización volumétrica(Fig 14), debido a su relevancia y aporte en trabajos previos. Cada corte fue redimensionado a una resolución de 512×512 píxeles.

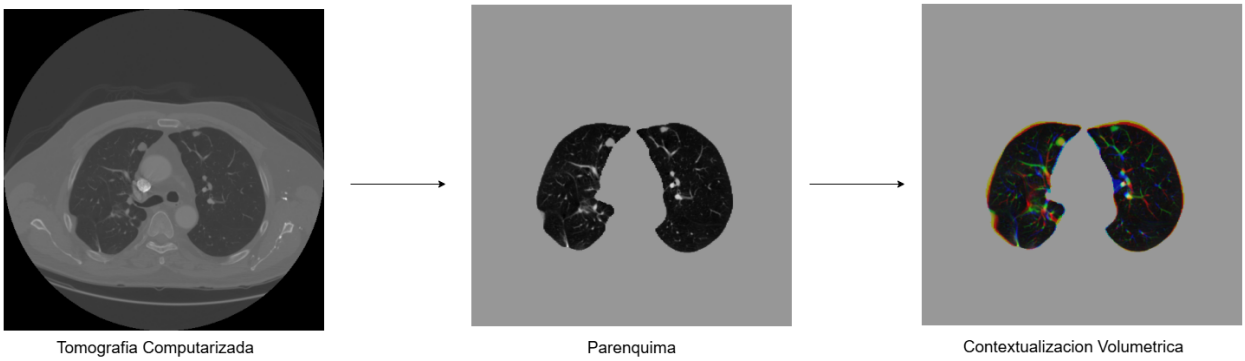

Figura 14. Procesamiento de contextualización volumétrica en secuencias TC. La imagen TC segmenta el parenquima pulmonar y se crea una imagen RGB a partir de 3 *slices* consecutivos

5.2. CONFIGURACIÓN DE LAS ARQUITECTURAS

Arquitectura fundacional. Para esta arquitectura se utilizó el modelo Grounding-DINO con parámetros preconfigurados. El proceso de aumento de datos empleó escalas de entrada que oscilaban entre 480 y 800 píxeles, con un límite máximo de 1333 píxeles. Las imágenes fueron redimensionadas utilizando escalas de 400, 500 y 600 píxeles, mientras que los recortes se realizaron en tamaños de 384 y 600 píxeles, cuidando que no se presentaran solapamientos entre ellos. El entrenamiento se ejecutó con un tamaño de lote de 3, utilizando el codificador de imágenes Swin-T 224 1k. Se incorporaron embebidos posicionales del tipo *sine*, configurados con una temperatura de 20. La arquitectura del *transformer* estuvo compuesta por 6 capas en el codificador y 6 en el decodificador, sin aplicar normalización previa. La dimensión del bloque *feedforward* fue de 2048, mientras que la de las capas ocultas se fijó en 256, sin utilizar *dropout*. Se configuraron 8 cabezas de atención y un total de 900 *queries* entrenables con dimensión 4. El número de puntos de atención para tanto el codificador como el decodificador fue de 4. Se empleó una estrategia estándar de dos etapas, sin compartir embebidos de las cajas delimitadoras para los candidatos a nódulos pulmonares (NP). La función de activación seleccionada para el *transformer* fue *ReLU*, y en el decodificador se compartieron los embebidos de las predicciones de las cajas. Como técnica de regularización, se introdujo ruido en las

cajas delimitadoras con una escala de 1.0 y una proporción de ruido en las etiquetas del 0.5. El texto de entrada se limitó a una longitud máxima de 256 caracteres, utilizando el codificador textual *bert-base-uncased*. La palabra clave principal empleada fue *"nodule"*, y se construyeron múltiples frases descriptivas como entrada al modelo, siguiendo una estructura semántica coherente. Las frases utilizadas fueron las siguientes:

- *a nodule in the lung*
- *a lung nodule with [lobulation-index], [spiculation-index], [subtlety-index]*
- *a pulmonary nodule described as [lobulation-index], [spiculation-index], [subtlety-index]*
- *a suspicious lung nodule showing [lobulation-index], [spiculation-index], [subtlety-index] traits*
- *a lung nodule characterized by [lobulation-index], [spiculation-index], [subtlety-index]*
- *a nodule located in the lung with [lobulation-index], [spiculation-index], [subtlety-index]*
- *a nodule with [lobulation-index]*
- *a nodule with [spiculation-index]*
- *a nodule with [subtlety-index]*

Donde los índices *[lobulation-index]*, *[spiculation-index]* y *[subtlety-index]* corresponden a los valores cuantitativos de los atributos de lobulación, espiculación y sutileza del nódulo. Estos valores fueron utilizados para personalizar cada frase, permitiendo que el modelo textual incorporara información morfológica relevante sobre cada nódulo pulmonar. El modelo fue optimizado con AdamW, con una tasa de aprendizaje base de 0.0001 y un parámetro de *weight decay* de 0.0001; tanto el codificador de texto como el de imágenes

se entrenaron con una tasa de 1e-05. El entrenamiento se ejecutó durante 30 épocas, reduciendo la tasa de aprendizaje a partir de la cuarta época y guardando el modelo con mejor rendimiento en la época correspondiente.

Red basada en atención. Para esta red se utilizó el modelo RT-DETR-L con parámetros preconfigurados. El aumento de datos empleó una resolución fija de 640 píxeles (sin multiescala), traslación (`translate=0.1`) y escalado (`scale=0.5`). El entrenamiento se ejecutó con un tamaño de lote de 12; se empleó el optimizador *AdamW*, `lr0=0.01` escalado a un valor efectivo de $\approx 0{,}001875$ tras 3 épocas de calentamiento, momento $0{,}937$ y decaimiento de pesos de 5×10^{-4}, durante 100 épocas.
La arquitectura del transformador estuvo compuesta por 6 capas en el descodificador deformable, con normalización por capa y sin abandono; la dimensión del bloque de alimentación fue de 1024 y la de las capas ocultas de 256. Se configuraron 8 cabezas de atención y $K = 300$ consultas entrenables con dimensión 4; se utilizaron embebidos posicionales sinusoidales. El número de puntos de atención por cabeza y nivel fue de 4 y se operó sobre 3 niveles de características. Se adoptó una estrategia estándar de dos etapas (propuestas del codificador y refinamiento en el descodificador) y la función de activación del transformador fue ReLU. En validación e inferencia no se aplicó supresión de no máximos (`nms=false`), dado que RT-DETR no requiere de dicha etapa.

Modelo reductor de FP. Esta estrategia utilizó el modelo *EfficientNet-B0*,inicializado con pesos preentrenados. Para la reducción de falsos positivos (FP) se empleó la arquitectura *EfficientNet-B0*, la cual fue adaptada para una tarea de clasificación binaria (FP vs. TP). El modelo fue inicializado con pesos preentrenados en *ImageNet* y su capa final totalmente conectada fue reemplazada por un bloque compuesto de una capa *dropout* de 0.5 y una capa densa con dos neuronas de salida. El conjunto de datos empleado se dividió en entrenamiento (80 %) y validación (20 %), manteniendo un balance entre clases. Para

incrementar la robustez del modelo y reducir el sobreajuste, se aplicaron técnicas de aumento de datos sobre las imágenes: recorte aleatorio redimensionado a 224×224, rotaciones de hasta $15°$, volteos horizontales y verticales, así como ajustes de brillo y contraste. El entrenamiento se realizó con un tamaño de lote de 16, una tasa de aprendizaje inicial de 1×10^{-4} y un total de 30 épocas. El optimizador seleccionado fue *Adam* con regularización L_2 (*weight decay* de 1×10^{-5}), y como función de pérdida se utilizó *Cross-Entropy Loss*. Durante el entrenamiento, se monitorizaron métricas de desempeño como *accuracy*, *loss* y el área bajo la curva ROC (AUC). El mejor modelo se guardó en función del valor más alto de AUC alcanzado en el conjunto de validación.

5.3. VALIDACIÓN

En cuanto a las métricas de validación, para las tareas de detección, en este trabajo se midió el desempeño del modelo cuantificando la cantidad de predicciones acertadas, la frecuencia de falsos positivos tanto por imagen como por estudio completo de TC *(scan)*, así como la precisión promedio de las detecciones. A continuación se describen en detalle las métricas utilizadas: **Precisión media promedio (mAP@IOU):** La métrica mAP@IoU (del inglés *mean average precision*) se emplea para medir la capacidad del modelo en la detección de NP con un umbral de Intersección sobre Unión (IoU), evaluando qué tan bien coinciden las predicciones con las anotaciones de referencia del conjunto de prueba. Esta métrica refleja la precisión del modelo al identificar NP cuya superposición con las etiquetas supera un umbral de IoU predefinido. El valor de mAP se obtiene promediando la precisión sobre todos los casos evaluados

$$mAP = \frac{1}{k} \sum_{i}^{k} AP_i, \; AP = \int_0^1 p(r) dr$$

donde AP_i representa la precisión media correspondiente a la clase i. Esta se determina como el área bajo la curva precisión-recall (precisión-sensibilidad), denotada como $p(r)$,

que cuantifica la relación entre la proporción de verdaderos positivos respecto al total de predicciones positivas (precisión) y la proporción de verdaderos positivos respecto al total de instancias reales (sensibilidad), a través de distintos umbrales de confianza. Una vez obtenidas las AP_i para todas las clases, su promedio define el valor de mAP. Un valor elevado de mAP indica un mejor desempeño del modelo, evidenciando una mayor precisión y sensibilidad en la detección de los objetos evaluados.

Métrica de Rendimiento de Competición (CPM): Esta métrica CPM (del inglés *Competition Performance Metric*) permite evaluar la sensibilidad del modelo en función de distintas cantidades de FP por imagen o por estudio completo (*scan*). Para calcularla, se varía el umbral de confianza de las predicciones, lo cual influye directamente en el número de predicciones consideradas como verdaderos positivos. Este ajuste genera diferentes pares de sensibilidad y tasa de FP. Posteriormente, se calcula la sensibilidad media promediando los valores correspondientes a un conjunto específico de niveles de FP, normalmente siete, lo que proporciona una visión más completa del desempeño del modelo bajo distintos niveles de tolerancia a falsos positivos. Formalmente, esto se expresa como $CPM = \frac{1}{7}\sum_{c=1}^{7} r(FP_c)$, donde c representa cada nivel predefinido de falsos positivos por imagen o *scan* (por ejemplo, 0.125, 0.25, 1, 2, 4, 8), y $r(FP_c)$ es la sensibilidad del modelo según el número de FP_c falsos positivos. Este enfoque permite una evaluación más completa del rendimiento del modelo, al mostrar su capacidad para mantener una alta sensibilidad mientras se limita la generación de falsos positivos.

Curva FROC: Esta curva permite representar gráficamente la relación obtenida en el CPM. La curva FROC (del inglés *Free-Response Receiver Operating Characteristic*) muestra cómo varía la sensibilidad del modelo frente a diferentes tasas de falsos positivos por imagen o por estudio completo (*scan*). Esta representación facilita la visualización del comportamiento del modelo a distintos umbrales de confianza, destacando la compensación inherente entre sensibilidad y la generación de FP.

Para evaluar el rendimiento de los modelos de detección de NP, se llevaron a cabo análisis

tanto a nivel de imagen como de *scan*. Para el modelo fundacional, se realizaron seis pruebas distintas, en las que se variaron las entradas textuales (*text prompts*) utilizadas durante la inferencia. Este procedimiento tuvo como objetivo examinar el aporte del texto sobre el rendimiento de detección. Cada prueba correspondió a una de las siguientes descripciones, las cuales varían en complejidad y especificidad semántica respecto a los rasgos visuales esperados en los nódulos pulmonares:

- *nodule*
- *a suspicious lung nodule showing no lobulation, no spiculation and obvious traits*
- *a suspicious lung nodule showing no lobulation and obvious traits*
- *a suspicious lung nodule showing no lobulation, no spiculation*
- *a suspicious lung nodule showing no spiculation*
- *a suspicious lung nodule showing obvious traits*

En cuanto al análisis a nivel de imagen, se trabajó únicamente con aquellas muestras del conjunto de prueba que contenían NP. Esta estrategia permitió enfocar la evaluación en la efectividad del modelo para identificar y localizar nódulos en contextos clínicamente relevantes, minimizando la influencia de imágenes sin hallazgos. Así, se obtuvo una medición del rendimiento del modelo en escenarios donde está asegurada la presencia de NP. Asimismo, se utilizó un umbral de *IoU* de 0.25 para la evaluación, dado que este valor es el más comúnmente empleado en tareas de detección de objetos pequeños como los nódulos pulmonares, permitiendo así una medida más realista del desempeño del modelo en este dominio específico.

6. EVALUACIÓN Y RESULTADOS

En este trabajo se implementaron dos arquitecturas contextuales para la detección de nódulos: un enfoque fundacional y un enfoque ligero, basado en *transformers*. Además, se diseñó una validación Profesor–Estudiante para aprovechar el aprendizaje de una arquitectura de mayor escala en una arquitectura relativamente ligera. Los enfoques propuestos fueron validados a nivel de imagen y a nivel de volumen (estudio de TC completo) y sobre dos conjuntos de datos distintos. En las siguientes subsecciones se detallan los resultados de estas evaluaciones.

6.1. MODELO FUNDACIONAL

6.1.1. Caracterización por imagen El modelo fundacional incluye tanto entrada de imágenes como de texto, lo cual puede afectar sus resultados. Para este trabajo fue de gran interés reconocer cómo dichas variaciones en el texto pueden afectar la detección de nódulos. De hecho, esta información es de vital importancia para transferir a la guía clínica información sobre cómo usar estos modelos y cómo hacer las consultas respectivas. Así, en el primer experimento se diseñaron sentencias textuales típicamente usadas por radiólogos durante el análisis de imágenes, para evaluar el comportamiento del modelo fundacional ajustado para esta tarea. Entonces, se evaluó el impacto de distintas entradas textuales (*prompts*) en el rendimiento del detector sobre el conjunto de datos LIDC-IDRI con el que fue ajustado. La comparación se realizó en términos de las métricas mAP y CPM, con el fin de analizar en qué medida el contenido del texto influye en la capacidad de localización de nódulos pulmonares, como se reporta en la Tabla 2.

De acuerdo con los resultados presentados en la Tabla 2, la entrada textual “nodule” obtuvo el mejor desempeño, alcanzando un *mAP* de 0.837 y un *CPM* de 0.851. Sin embargo, los resultados no tienen un impacto significativo con respecto a las demás variantes de

Prompt	CPM	mAP
nodule	**0.851 ± 0.009**	**0.837**
suspicious lung nodule	0.846 ± 0.003	0.833
a suspicious lung nodule showing no lobulation, no spiculation and obvious traits	0.830 ± 0.005	0.816
a suspicious lung nodule showing no lobulation and obvious traits	0.839 ± 0.006	0.824
a suspicious lung nodule showing no lobulation, no spiculation	0.831 ± 0.001	0.816
a suspicious lung nodule showing no spiculation	0.833 ± 0.003	0.819

Tabla 2. Resultados comparativos del modelo fundacional bajo diferentes *prompts* de texto. Se muestran los valores de CPM para intervalos de 1/8 a 8, así como los valores de mAP.

entrada. Lo anterior puede sugerir que el modelo fundacional mantiene un rendimiento sólido en la tarea de detección, independientemente de la formulación textual utilizada. Estos resultados también pueden sugerir exploraciones más profundas en nuevos trabajos para aprovechar la información textual resultante. Además, cabe resaltar los resultados sobresalientes en detección logrados por el modelo fundacional, lo que marca una ventaja clara al haber aprovechado su base de entrenamiento sobre un gran número de datos, en diferentes tareas, logrando una representación robusta con un ajuste relativamente sencillo. Considerando este carácter de generalización, en un segundo experimento se decidió evaluar el modelo en el conjunto de datos NLST, el cual contiene estudios de TC con un carácter diferente en la representación visual. Para ello, se obtuvieron las curvas de *Precision* y *Recall*. La Figura 15 resume los resultados obtenidos en cuanto a las estimaciones del modelo fundacional en este conjunto de datos externo. Bajo la entrada textual “nodule”, los resultados en el conjunto NLST muestran un *mAP* de apenas 0.273 (Figura 15-a) y un *CPM* de 0.396 (Figura 15-b). Estos valores reflejan las limitaciones del modelo al aplicarse en datos externos, evidenciando una pérdida significativa de rendimiento fuera del dominio de entrenamiento.

6.1.2. Caracterización por volumen En una validación adicional, en este trabajo también se consideró la evaluación de los resultados desde el origen volumétrico de los estudios de TC. En este análisis se evaluó el impacto de distintas entradas textuales en

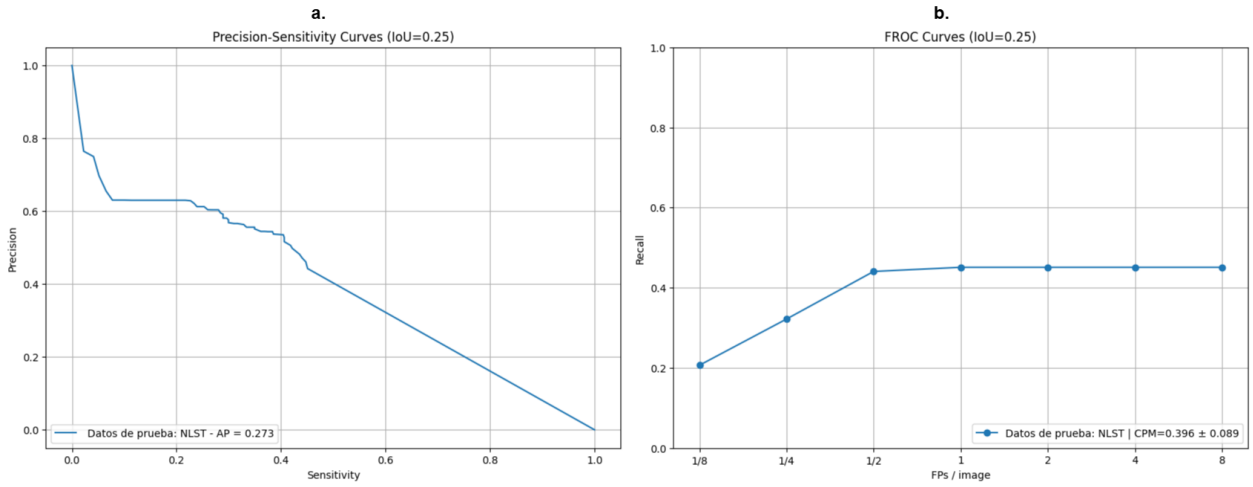

Figura 15. Resultados del modelo fundacional sobre el conjunto de datos NLST: (a) resultados en términos de mAP y (b) resultados en términos de CPM.

el rendimiento del detector sobre el conjunto de datos LIDC-IDRI, considerando tanto los resultados directos del modelo como aquellos obtenidos tras la aplicación del RFP (reductor de falsos positivos). La comparación se realizó en términos de *recall* para diferentes números de falsos positivos por *scan* (FP/*scan*) y de la métrica CPM, con el propósito de analizar en qué medida la formulación del texto influye en la capacidad de localización de NP a nivel de volumen. Es interesante que, en esta evaluación, se observa que la representación textual "*a suspicious lung nodule showing no lobulation, no spiculation*" alcanzó el mejor desempeño global en términos de CPM, con un valor de 0.469. Sin embargo, las demás entradas textuales también mostraron un rendimiento sólido, manteniendo valores consistentes tanto en CPM como en *recall*. Por su parte, la entrada "nodule" destaca porque alcanzó el mayor valor de *recall* (0.904) con 8 FP/*scan* tras la aplicación del RFP, lo que evidencia que incluso una descripción simple puede ofrecer un alto nivel de sensibilidad en la detección de nódulos pulmonares.

En la misma línea de acción desarrollada en el análisis por imágenes, en este análisis volumétrico también se evaluó el rendimiento del modelo fundacional en el conjunto de datos NLST empleando la entrada textual *"nodule"*. En la Figura 16-a, la curva FROC

Prompt	Método	1/8	1/4	1/2	1	2	4	8	CPM
nodule	BASE	0.081	0.081	**0.274**	0.343	0.542	0.705	0.860	0.413 ± 0.279
	RFP	0.079	0.079	0.270	0.411	0.612	0.815	**0.904**	0.453 ± 0.310
suspicious lung nodule	BASE	0.079	0.079	0.268	0.337	0.530	0.705	0.856	0.408 ± 0.279
	RFP	0.077	0.077	0.264	0.392	0.593	0.819	0.892	0.445 ± 0.309
a suspicios lung nodule showing no lobulation, no spiculation and obvious traits	BASE	**0.091**	0.124	0.211	0.335	0.518	0.744	0.876	0.414 ± 0.285
	RFP	**0.091**	0.166	0.203	0.404	**0.642**	**0.827**	0.880	0.459 ± 0.301
a suspicious lung nodule showing no lobulation and obvious traits	BASE	0.087	0.130	0.258	0.343	0.508	0.720	0.872	0.417 ± 0.275
	RFP	0.085	0.128	0.248	0.396	0.616	0.821	0.886	0.454 ± 0.301
a suspicious lung nodule showing no lobulation, no spiculation	BASE	0.089	0.138	0.238	0.380	0.526	0.730	0.876	0.425 ± 0.277
	RFP	0.089	**0.189**	**0.274**	**0.419**	0.610	0.825	0.876	$\mathbf{0.469 \pm 0.287}$
a suspicious lung nodule showing no spiculation	BASE	0.061	0.132	0.252	0.348	0.563	0.738	0.872	0.424 ± 0.286
	RFP	0.059	0.132	0.246	0.410	0.616	**0.827**	0.874	0.452 ± 0.304

Tabla 3. Resultados comparativos del modelo fundacional y reductor de falsos positivos bajo diferentes *prompts* de texto. Se muestran los valores de recall para FP de 1/8 a 8, así como el CPM.

correspondiente al intervalo de 1/8 a 8 FP/*scan* muestra un CPM de apenas 0.142 para el modelo base, mientras que, con el reductor de falsos positivos, este valor cae drásticamente hasta 0.009. Este comportamiento evidencia que el modelo presenta serias dificultades para mantener la sensibilidad en un escenario de bajas tasas de falsos positivos, lo que sugiere una mayor incidencia de falsos negativos al enfrentarse a este nuevo dominio de datos. De manera similar, en la Figura 16-b, para el rango de 4 a 256 FP/*scan*, el modelo base alcanza un CPM de 0.420 y el modelo con reductor un valor de 0.389. Aunque el desempeño mejora respecto al escenario anterior, la diferencia entre ambos enfoques sigue siendo reducida y los valores continúan por debajo de lo obtenido en LIDC. En conjunto, estos resultados reflejan que el modelo no logra adaptarse de manera satisfactoria a NLST, lo que confirma limitaciones en su capacidad de generalización y, especialmente, en la detección de nódulos en contextos distintos al del ajuste original, donde los falsos negativos juegan un papel determinante.

6.1.3. Análisis de características radiológicas de NP Una vez realizada la evaluación a nivel de imagen y volumen de los modelos implementados, así como la inclusión del reductor de falsos positivos (RFP), en esta sección se presenta un análisis complementario

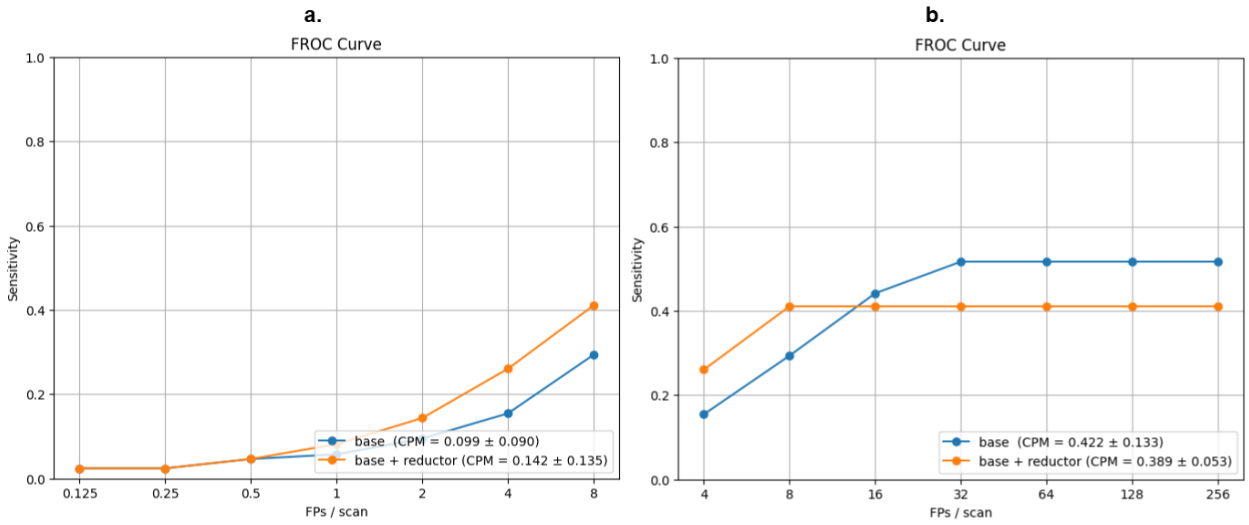

Figura 16. Resultados por volumen del modelo fundacional sobre el conjunto de datos NLST: (a) curva FROC en el rango de 1/8 a 8 FP/scan y (b) curva FROC en el rango de 4 a 256 FP/scan. En ambos casos se muestran los valores de CPM para el modelo base y con reductor de falsos positivos.

que examina la capacidad de detección del modelo fundacional según características morfológicas de los nódulos pulmonares, como lobulación, espiculación y sutileza (ver Figura 17). Este análisis permite identificar fortalezas y limitaciones específicas del modelo en la detección de distintos tipos de NP, utilizando de forma estratificada las representaciones textuales incorporadas en este trabajo, lo que facilita una interpretación más clara de los resultados desde una perspectiva clínica.

Como se muestra en la figura 17-a, en el análisis estratificado por lobulación el modelo alcanzó una tasa de detección del 83.5 % para NP sin lobulación, 90.7 % para aquellos con lobulación mínima, 82.4 % para nódulos con lobulación moderada, 91.3 % con lobulación casi marcada y 82.4 % en NP con lobulación marcada. En la evaluación estratificada por espiculación (figura 17-b), el modelo mostró una tasa de detección del 84.5 % para NP sin espiculación, 89.4 % en aquellos con espiculación mínima, 82.1 % para nódulos con espiculación moderada, 88.2 % en los de espiculación casi marcada y 86.4 % en NP con espiculación marcada. Estos resultados evidencian un desempeño sólido del modelo frente

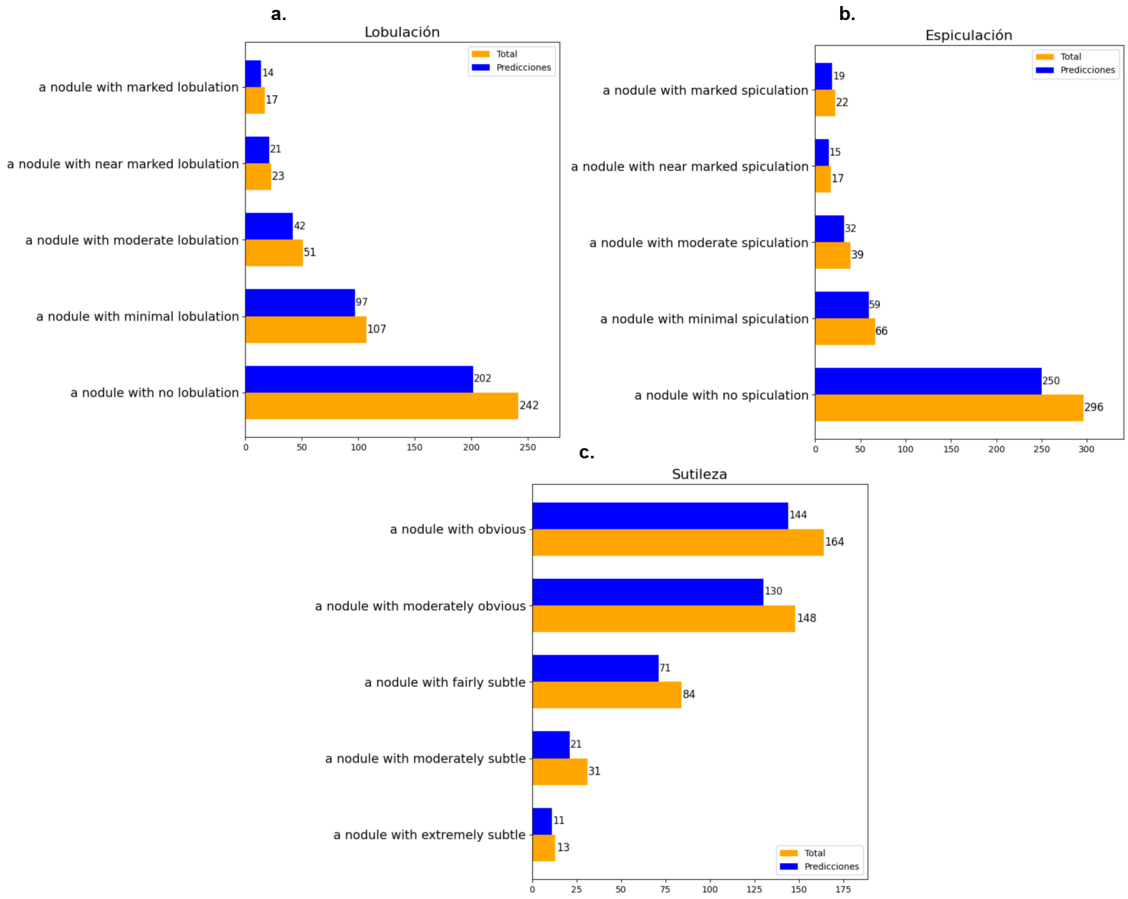

Figura 17. Modelo fundacional: Resultados de caracterización de los NP, a) diagrama de barras correspondiente al desempeño por lobulación, b) por espiculación y c) por sutileza o dificultad de detección según los radiólogos

a la detección de nódulos con contornos irregulares, lo que resulta relevante dado que la lobulación y la espiculación son rasgos asociados al grado de malignidad o benignidad de un nódulo. En la evaluación por sutileza, véase la figura 17-c, el modelo mostró una tasa de detección para los NP extremadamente sutiles de 84.6 %, seguida de 67.7 % en aquellos que poseen una sutileza moderada, 84.5 % para nódulos con sutileza, 87.8 % para los NP moderadamente obvios y 87.8 % en NP extremadamente obvios. Este resultado evidencia que el modelo es capaz de reconocer incluso los nódulos pulmonares más complejos, alcanzando tasas de detección superiores al 67 % en los casos de sutileza extrema y moderada. Este hallazgo adquiere gran relevancia en el ámbito clínico, ya que

precisamente los nódulos más difíciles de identificar son los que suelen pasar inadvertidos en la práctica habitual, lo que repercute de manera directa en el diagnóstico temprano.

6.2. MODELO BASADO EN ATENCIÓN

En cuanto al modelo basado en bloques de atención (*RT-DETR*), en este trabajo se entrenó exclusivamente con anotaciones del conjunto LIDC. A continuación se detallan los resultados obtenidos tanto sobre LIDC como sobre NLST, tanto por imagen como evaluando el volumen de TC.

6.2.1. Caracterización por imagen La Tabla 4 reporta los resultados obtenidos en los dos conjuntos de datos, en términos de las métricas definidas en este trabajo: *CPM* y *mAP*. La evaluación se realizó frente a un conjunto de validación del conjunto de datos LIDC. Cabe recordar que una partición de este *dataset* fue usada en el entrenamiento. También se validó frente a NLST, pero esta vez sin ningún tipo de ajuste adicional, solo tomando los datos como inferencia.

Datos de prueba	CPM	mAP
LIDC	0.855 ± 0.038	0.821
NLST	0.388 ± 0.099	0.255

Tabla 4. Resultados comparativos del modelo RT-DETR entrenado con anotaciones reales de LIDC. Se muestran los valores de *CPM* y *mAP* sobre los conjuntos de prueba LIDC y NLST.

En el conjunto de datos LIDC, cuando el modelo fue entrenado y validado dentro del mismo dominio, alcanzó un desempeño destacado, con un *CPM* de 0.855 ± 0.038 y un *mAP* de 0.821. En contraste, al entrenar en LIDC y validar en NLST, los resultados globales fueron considerablemente más bajos, con un *CPM* de 0.388 ± 0.099 y un *mAP* de 0.255. Este comportamiento era esperable, dado que NLST corresponde a una distribución distinta de la utilizada en entrenamiento. Esto refleja la dificultad inherente al problema de generalización en aprendizaje profundo: los modelos tienden a ajustarse a las características

específicas del conjunto de entrenamiento (protocolos de adquisición, tipos de escáneres, características demográficas de los pacientes), por lo que su rendimiento se degrada al enfrentarse a dominios con distribuciones diferentes. A pesar de estos resultados, la arquitectura también resulta favorable cuando existen anotaciones que pueden considerarse inicialmente para ajustar el dominio durante el entrenamiento, permitiendo así lograr un comportamiento favorable.

6.2.2. Caracterización por volumen En cuanto a la validación volumétrica, el modelo basado en atención fue evaluado considerando todas las imágenes del estudio de TC, con el objetivo de emular su uso como herramienta de apoyo al diagnóstico en un escenario más cercano a la práctica clínica real. La evaluación se llevó a cabo en términos de sensibilidad (*recall*) para diferentes niveles de falsos positivos por *scan* (FP/*scan*) ($\frac{1}{8}, \frac{1}{4} \ldots 8$) y de la métrica CPM, considerando tanto los resultados base como aquellos obtenidos tras la incorporación del reductor de falsos positivos (RFP). La Tabla 5 reporta los resultados obtenidos en esta evaluación, tanto para el modelo de atención en forma cruda como después del RFP.

Datos de prueba	Método	1/8	1/4	1/2	1	2	4	8	CPM
LIDC	BASE	**0.069**	**0.132**	0.242	0.384	0.606	0.768	0.886	0.441 ± 0.295
	RFP	0.063	0.111	**0.289**	**0.486**	**0.671**	**0.819**	0.894	$\mathbf{0.476 \pm 0.309}$
NLST	BASE	0.015	0.026	0.066	0.117	0.170	0.260	0.364	0.146 ± 0.119
	RFP	**0.018**	**0.038**	**0.072**	**0.132**	**0.194**	**0.316**	**0.423**	$\mathbf{0.171 \pm 0.140}$

Tabla 5. Resultados a nivel de volumen del modelo RT-DETR entrenado con LIDC, con y sin reductor de falsos positivos (RFP).

En el conjunto de datos LIDC, el modelo entrenado con anotaciones realizadas por radiólogos alcanzó un desempeño consistente, con un *CPM* de 0.441 en el escenario base y de 0.476 al incorporar el reductor de falsos positivos (RFP). Estos valores reflejan la capacidad del modelo para identificar nódulos pulmonares de manera más robusta cuando se aplican estrategias adicionales de reducción de falsos positivos dentro del

mismo dominio de entrenamiento. En el conjunto de datos NLST se obtuvo un *CPM* de 0.146 en el escenario base y de 0.171 al aplicar el RFP. Este comportamiento era esperable, dado que NLST presenta una mayor heterogeneidad en la calidad de las imágenes, lo que dificulta la capacidad de generalización del modelo entrenado con LIDC hacia este dominio. Este resultado resulta interesante para comprender que, en entornos heterogéneos como los persistentes en la clínica, se requiere un nivel de ajuste de los modelos a los dispositivos de captura, los protocolos de anotación e incluso a los acuerdos y experiencias establecidas en cada centro hospitalario. Sin embargo, este ajuste no es necesariamente realista en entornos hospitalarios con limitaciones de cómputo, por lo cual se abre la posibilidad de explorar nuevas técnicas de entrenamiento que permitan transferir y ajustar las arquitecturas a distribuciones de imágenes, pero sin requerimiento de supervisión.

6.3. DESTILACIÓN DE CONOCIMIENTO

Como exploración adicional y soportada en los resultados obtenidos, en este trabajo se incluyó la exploración inicial de un esquema de destilación de conocimiento desde una arquitectura fundacional hacia una arquitectura más compacta RT-DETR. El experimento consistió en simular un escenario en el que no existen etiquetas para ajustar un modelo, considerando condiciones de nuestros centros hospitalarios, *i.e.*, sin radiólogos expertos para generar las anotaciones. Entonces, se utilizó una arquitectura con mayor confianza y capacidad de generalización para generar las etiquetas, las cuales fueron usadas por el modelo compacto para ajustar su representación a nuevas observaciones. Así, en esta sección se presentan los resultados obtenidos al entrenar el modelo RT-DETR con pseudo-etiquetas generadas sobre el conjunto NLST mediante el modelo fundacional. Este enfoque busca explorar el potencial de la estrategia de destilación Profesor–Estudiante, en la que un modelo de gran capacidad proporciona supervisión automática para entrenar un modelo más ligero.

6.3.1. Caracterización por imagen Para el modelo basado en atención (*RT-DETR*) entrenado con pseudo-etiquetas, se evaluó en los dos conjuntos de datos previamente mencionados, con el fin de analizar el impacto que tiene la fuente de anotación en el desempeño del modelo. Los resultados comparativos en términos de *CPM* y *mAP* se presentan en la Tabla 6.

Datos de prueba	CPM	mAP
LIDC	0.810 ± 0.035	0.787
NLST	0.424 ± 0.048	0.349

Tabla 6. Resultados comparativos del modelo RT-DETR entrenado con pseudo-etiquetas generadas en NLST. Se muestran los valores de *CPM* y *mAP* sobre los conjuntos de prueba LIDC y NLST.

En el conjunto de datos LIDC, el modelo entrenado a partir de pseudo-etiquetas generadas automáticamente alcanzó un desempeño competitivo, con un *CPM* de 0.810 ± 0.035 y un *mAP* de 0.787. Estos valores reflejan que, aún sin contar con anotaciones manuales realizadas por especialistas, el modelo fue capaz de aprender representaciones efectivas de los nódulos pulmonares a partir de etiquetas derivadas de un modelo fundacional. En el conjunto de datos NLST, los resultados globales no fueron tan elevados como en LIDC, con un *CPM* de 0.424 ± 0.048 y un *mAP* de 0.349. Este comportamiento era esperable, dado que NLST presenta una mayor variabilidad en la calidad de las imágenes y no dispone de anotaciones realizadas directamente por radiólogos. No obstante, el uso de pseudo-etiquetas derivadas de este mismo dominio permitió al modelo adaptarse mejor a sus características específicas, mostrando un desempeño relativamente superior frente a la validación cruzada con LIDC y evidenciando el potencial de la estrategia de pseudo-etiquetado para mejorar la generalización en escenarios con disponibilidad limitada de anotaciones manuales.

6.3.2. Caracterización por volumen En la misma línea de validación volumétrica, en esta exploración Profesor–Estudiante, se evaluó el desempeño del modelo *RT-DETR* a

partir de las etiquetas generadas por el modelo fundacional. El análisis, en términos de *recall* y *CPM*, incluyó tanto los resultados base como los obtenidos tras aplicar el RFP. De esta manera, se buscó determinar en qué medida la estrategia de anotación y la aplicación del RFP influyen en la capacidad del modelo para detectar nódulos pulmonares en el escenario volumétrico. La Tabla 7 reporta los resultados obtenidos en este esquema de entrenamiento.

Datos de prueba	Método	1/8	1/4	1/2	1	2	4	8	CPM
LIDC	BASE	**0.026**	**0.067**	**0.136**	0.238	0.453	0.711	0.886	0.360 ± 0.310
	RFP	**0.026**	0.063	0.118	**0.264**	**0.530**	**0.797**	**0.911**	$\mathbf{0.387 \pm 0.335}$
NLST	BASE	**0.115**	**0.161**	**0.194**	0.245	**0.331**	0.399	**0.481**	$\mathbf{0.275 \pm 0.124}$
	RFP	0.097	0.139	0.172	**0.258**	0.320	**0.404**	0.450	0.263 ± 0.125

Tabla 7. Resultados a nivel de volumen del modelo RT-DETR entrenado con pseudo-etiquetas en NLST, con y sin RFP.

El modelo entrenado con pseudo-etiquetas alcanzó un *CPM* inferior (0.360 en el escenario base y 0.387 con la incorporación del RFP). Sin embargo, en este último caso se registró la mayor sensibilidad observada, con un valor de 0.911 en el nivel de 8 FP/*scan*. Este comportamiento indica que, si bien la estrategia de pseudo-etiquetas combinada con RFP maximiza la capacidad de detección en umbrales permisivos de falsos positivos, lo hace sacrificando desempeño en escenarios más restrictivos, donde se incrementa el número de falsos negativos. En el conjunto de datos NLST, los resultados a nivel de volumen tampoco alcanzaron valores tan elevados como en LIDC, con un *CPM* de 0.275 en el escenario base y 0.263 al incorporar el RFP. Este comportamiento era esperable, dado que la validación volumétrica introduce un mayor nivel de complejidad y sensibilidad a la variabilidad interescáner, lo que dificulta la generalización del modelo. No obstante, el hecho de que estas métricas provengan de pseudo-etiquetas generadas dentro del mismo dominio sugiere una mejor adaptación del modelo a las características particulares de NLST, replicando la tendencia observada en el análisis por imagen y reforzando el valor del pseudo-etiquetado como estrategia para afrontar la escasez de anotaciones manuales

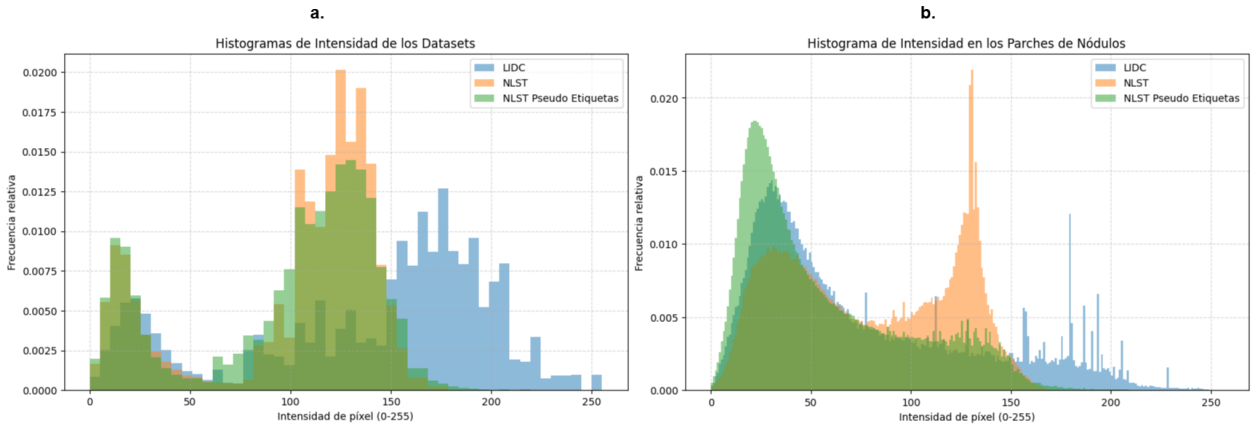

Figura 18. Distribución de intensidades de píxeles en (a) imágenes completas y (b) parches que contienen nódulo o candidatos a nódulo pulmonares, para los distintos conjuntos de datos. El histograma azul corresponde a etiquetas reales del LIDC, el histograma naranja a etiquetas reales del NLST y el histograma verde a pseudo-etiquetas generados sobre el NLST.

en contextos clínicos reales.

6.4. ANÁLISIS DE LOS CONJUNTOS DE DATOS

Considerando los resultados obtenidos en las secciones anteriores, tanto para el modelo fundacional como para el modelo de atención (entrenado con anotaciones reales y con pseudo-etiquetas), se evidenció una marcada discrepancia en el desempeño al validar en diferentes conjuntos de datos. Por ejemplo, a nivel de imagen en LIDC se alcanzaron valores de *CPM* de 0.855 (RT-DETR con LIDC), 0.851 (Grounding DINO) y 0.810 (RT-DETR con pseudo-etiquetas), mientras que en NLST los resultados fueron considerablemente más bajos, con 0.388, 0.396 y 0.424, respectivamente. Esta diferencia motivó la realización de un análisis adicional sobre las características propias de cada conjunto, particularmente la distribución de intensidades tanto en las imágenes completas como en los parches correspondientes a nódulos pulmonares, con el fin de comprender los factores que podrían estar influyendo en dicha variabilidad de rendimiento.

Como se aprecia en la Figura 18-a, las imágenes completas del conjunto LIDC presentan una distribución de intensidades claramente diferenciada respecto al conjunto NLST, así

como frente al conjunto derivado de pseudo-etiquetas. En este último caso, la similitud respecto a NLST se explica porque las imágenes base provienen del mismo conjunto de datos. Sin embargo, en la Figura 18-b se observa que los parches centrados en NP siguen una tendencia distinta: aunque las distribuciones de intensidades de LIDC y NLST mantienen diferencias notorias, los NP obtenidos a partir de pseudo-etiquetas muestran un comportamiento más cercano al de LIDC que al del propio NLST. Esto se debe a que las regiones etiquetadas fueron generadas por el modelo fundacional, previamente entrenado en LIDC, lo cual introduce un sesgo hacia la distribución de dicho conjunto. Por ejemplo, en el rango de intensidades comprendido entre 50 y 150, las distribuciones de LIDC y pseudo-etiquetas presentan similitudes claras entre sí, pero se apartan significativamente de la distribución original de NLST.

Estas observaciones sugieren que la brecha de rendimiento entre LIDC y NLST tiene una relación directa con las diferencias estadísticas de los dominios, pues los modelos entrenados en LIDC fueron validados sobre un conjunto con características de intensidad y representación distintas. De manera similar, el modelo entrenado con pseudo-etiquetas heredó el sesgo de LIDC en la selección de parches, lo que explica que, aunque haya alcanzado valores elevados de *CPM* en LIDC, su desempeño se redujo de forma considerable en NLST. No obstante, la implementación de un método de destilación de conocimiento basado en pseudo-etiquetas permitió observar leves mejoras: a nivel de imagen, el *CPM* pasó de 0.388 y un *mAP* de 0.255 (modelo entrenado directamente en LIDC) a 0.424 y 0.349, respectivamente; mientras que a nivel de volumen, se pasó de 0.146 ± 0.119 a 0.275 ± 0.124 sobre el conjunto NLST. Estos resultados sugieren que métodos de destilación de conocimiento más robustos podrían favorecer la adaptación de modelos de detección a diferentes dominios de TC.

7. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo presentó dos arquitecturas orientadas a capturar las relaciones contextuales de los nódulos pulmonares (NP) en imágenes de tomografía computarizada (TC). La primera corresponde a una arquitectura fundacional, Grounding-DINO, que integró las características radiológicas de los nódulos pulmonares como contexto semántico; y la segunda se basa en mecanismos de atención (*transformer*), específicamente RT-DETR. Adicionalmente, se implementó una estrategia de destilación de conocimiento (Profesor–Estudiante), transfiriendo la información aprendida por la arquitectura fundacional hacia el modelo tipo *transformer*, con el objetivo de abordar el problema en contextos caracterizados por restricciones de infraestructura computacional. Tanto el modelo fundacional como el modelo *transformer* fueron entrenados utilizando el conjunto de datos LIDC-IDRI. En contraste, bajo la estrategia de Profesor–Estudiante, el modelo fundacional generó pseudo-etiquetas sobre el conjunto NLST, el cual carece de anotaciones precisas de la ubicación de los nódulos pulmonares, salvo en 88 de los 2.086 estudios disponibles. Todas las estrategias fueron validadas tanto en el conjunto LIDC como en NLST, empleando únicamente los *scans* con anotaciones de referencia. Sobre el conjunto LIDC, a nivel de imagen, el modelo fundacional alcanzó su mejor desempeño con un *CPM* de 0.851 al utilizar el *prompt* “nodule”. Adicionalmente, el modelo mostró un desempeño destacado en características radiológicas que dificultan la detección de nódulos. Detectó el 84.6 % de los nódulos extremadamente sutiles, considerados uno de los principales desafíos para los radiólogos debido a su bajo contraste y reducido tamaño, y mantuvo un buen rendimiento en casos con lobulación y espiculación. De esta forma, los resultados obtenidos evidencian robustez ante distintos patrones morfológicos. Por otra parte, el modelo *transformer* alcanzó un CPM de 0.855, mientras que la estrategia Profesor–Estudiante logró 0.810, lo que demuestra la efectividad del entrenamiento supervisado directo. Sin embargo, también resalta el potencial del pseudo-etiquetado para obtener resultados competitivos

en escenarios con disponibilidad limitada de anotaciones manuales. Estos resultados, en términos de CPM, superan y se mantienen competitivos frente a reportes del estado del arte, que alcanzan valores de 0.79 [54] y 0.88 [55], respectivamente. Cabe resaltar que el presente trabajo evaluó directamente sobre el conjunto completo LIDC, mientras que los estudios comparados emplearon LUNA16 [56], un subconjunto del mismo que excluye nódulos pequeños y casos de baja sutileza. A nivel volumétrico, se implementó un método básico de reducción de falsos positivos mediante *EfficientNet*, con el fin de mitigar la alta proporción de detecciones erróneas. En el caso del modelo fundacional, se evaluaron distintas entradas textuales, lo que aportó robustez frente a diferentes descripciones. La mejor configuración se obtuvo con el *prompt* “a suspicious lung nodule showing no lobulation, no spiculation”, alcanzando un CPM de 0.469, mientras que con el *prompt* genérico “nodule” se logró un CPM de 0.453. Por su parte, el modelo *transformer* alcanzó un CPM de 0.476, evidenciando el mejor desempeño dentro de las estrategias evaluadas, seguido por la estrategia Profesor–Estudiante, que obtuvo un CPM de 0.387. Estos resultados demuestran que tanto el modelo de atención como el fundacional se mantienen competitivos frente a valores reportados en el estado del arte, donde los métodos basados en reductores de falsos positivos alcanzan CPM de 0.405, 0.524, 0.506 y 0.508 en configuraciones 2D [7], considerando además que dichos trabajos se entrenaron sobre LUNA16, un subconjunto de LIDC. Finalmente, la estrategia de destilación Profesor–Estudiante logró un CPM de 0.387, mostrando también un desempeño competitivo al entrenar sobre pseudo-etiquetas

---

[54] Hao TANG; Daniel R. KIM y Xiaohui XIE. «Automated pulmonary nodule detection using 3D deep convolutional neural networks». En: *arXiv (Cornell University)* (ene. de 2019). DOI: 10.48550/arxiv.1903.09876.

[55] Xia HUANG, *et al.* «Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks». En: *Computerized Medical Imaging and Graphics* 74 (mar. de 2019). DOI: 10.1016/j.compmedimag.2019.02.003.

[56] Arnaud Arindra Adiyoso SETIO, *et al.* «Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge». En: *Medical Image Analysis* 42 (jul. de 2017). DOI: 10.1016/j.media.2017.06.015.

generadas en el propio dominio NLST. En el conjunto de datos NLST, los modelos fundacional, *transformer* y la estrategia Profesor–Estudiante alcanzaron valores de CPM de 0.396, 0.388 y 0.424, respectivamente. Estos valores evidencian las limitaciones de los modelos cuando son entrenados en un dominio distinto, donde las diferencias en la distribución de intensidades de los píxeles entre conjuntos de datos explican, en parte, la disminución en el rendimiento. No obstante, la estrategia Profesor–Estudiante mostró mejoras consistentes tanto a nivel de imagen como de volumen, lo que resalta su potencial para favorecer la adaptación al dominio de prueba. Este hallazgo motiva la exploración futura de mecanismos de destilación de conocimiento más avanzados, capaces de favorecer una generalización más robusta y efectiva a través de distintos conjuntos de datos. Como trabajo futuro, se considera relevante profundizar en la evaluación de estrategias de destilación de conocimiento que permitan una mejor adaptación a diferentes dominios de datos, particularmente aquellos con mayor variabilidad y heterogeneidad. En este sentido, explorar mecanismos más sofisticados de transferencia entre modelos puede potenciar la estrategia Profesor–Estudiante presentada en este trabajo, facilitando que el conocimiento adquirido en un dominio sea aprovechado de manera más efectiva en contextos distintos. De igual manera, resulta pertinente avanzar en el enriquecimiento del modelo fundacional mediante la incorporación de un vocabulario especializado en el ámbito médico. En este sentido, el uso de modelos de lenguaje de gran escala (LLM) entrenados para el contexto clínico podría aportar descripciones textuales más precisas y alineadas con la práctica radiológica. Esto permitiría no solo ampliar la variabilidad de los *prompts* empleados, sino también integrar de manera más sistemática otras características radiológicas que favorezcan una representación contextual más robusta y, en consecuencia, una detección más eficaz de los NP.

BIBLIOGRAFÍA

AGNES, S Akila; ANITHA, J y SOLOMON, A Arun. «Two-stage lung nodule detection framework using enhanced UNet and convolutional LSTM networks in CT images». En: *Computers in Biology and Medicine* 149 (sep. de 2022), pág. 106059. DOI: 10.1016/j.compbiomed.2022.106059 (vid. pág. 28).

ALI, Zeeshan; IRTAZA, Aun y MAQSOOD, Muazzam. «An efficient U-Net framework for lung nodule detection using densely connected dilated convolutions». En: *The Journal of Supercomputing* 78.2 (jun. de 2021), págs. 1602-1623. DOI: 10.1007/s11227-021-03845-x (vid. pág. 30).

ARMATO III, Samuel G, *et al.* «The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans». En: *Medical physics* 38.2 (2011), págs. 915-931 (vid. págs. 16, 55).

AUBARD, Martin, *et al.* «Knowledge distillation in YOLOX-ViT for side-scan sonar object detection». En: *arXiv preprint arXiv:2403.09313* (2024) (vid. pág. 51).

BOMMASANI, Rishi, *et al. On the Opportunities and Risks of Foundation Models*. Ago. de 2021 (vid. pág. 23).

CAO, Haichao , *et al.* «A Two-Stage Convolutional Neural Networks for Lung Nodule Detection». En: *IEEE Journal of Biomedical and Health Informatics* (ene. de 2020). DOI: 10.1109/jbhi.2019.2963720 (vid. págs. 13, 77).

CARION, Nicolas, *et al.* «End-to-end object detection with transformers». En: *European conference on computer vision*. Springer. 2020, págs. 213-229 (vid. pág. 19).

CHENG, Tianheng, *et al.* «YOLO-World: Real-Time Open-Vocabulary Object Detection». En: *arXiv (Cornell University)* (ene. de 2024). DOI: 10.48550/arxiv.2401.17270 (vid. págs. 23, 26).

CHOROWSKI, Jan K, *et al.* «Attention-based models for speech recognition». En: *Advances in neural information processing systems* 28 (2015) (vid. pág. 18).

CIELLO, Annemilia Del, *et al.* «Missed lung cancer: when, where, and why?» En: *Diagnostic and Interventional Radiology* 23.2 (2017), págs. 118-126. DOI: 10.5152/dir.2016.16187 (vid. págs. 12, 33).

DEVLIN, Jacob, *et al.* «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *arXiv (Cornell University)* (ene. de 2018). DOI: 10.48550/arxiv.1810.04805 (vid. págs. 23, 25, 43).

DING, J., *et al.* «Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks». En: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2017, págs. 559-567 (vid. pág. 13).

DOSOVITSKIY, Alexey, *et al.* «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: *arXiv preprint arXiv:2010.11929* (2020) (vid. págs. 13, 19).

DUMA, Narjust; SANTANA-DAVILA, Rafael y MOLINA, Julian R. «Non–Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment». En: *Mayo Clinic Proceedings* 94.8 (2019), págs. 1623-1640. DOI: 10.1016/j.mayocp.2019.01.013 (vid. págs. 12, 33).

FERLAY, J., *et al.* «Cancer statistics for the year 2020: An overview». En: *International Journal of Cancer* 149.4 (2021), págs. 778-789 (vid. pág. 33).

GUO, Zhitao , *et al.* «Msanet: multiscale aggregation network integrating spatial and channel information for lung nodule detection». En: *IEEE Journal of Biomedical and Health Informatics* 26.6 (2021), págs. 2547-2558 (vid. pág. 31).

HAN, Liying, *et al.* «BiRPN-YOLOvX: A weighted bidirectional recursive feature pyramid algorithm for lung nodule detection». En: *Journal of X-Ray Science and Technology* 31.2 (ene. de 2023), págs. 301-317. DOI: 10.3233/xst-221310 (vid. pág. 30).

HAO, R.; QIANG, Y.; YAN, X., *et al.* «Juxta-vascular pulmonary nodule segmentation in PET-CT imaging based on an LBF active contour model with information entropy and joint vector». En: *Computational and Mathematical Methods in Medicine* 2018 (2018) (vid. pág. 16).

HU, Tianjiao, *et al.* «A lung nodule segmentation model based on the transformer with multiple thresholds and coordinate attention». En: *Scientific Reports* 14.1 (2024), pág. 31743 (vid. pág. 31).

HUANG, Xia, *et al.* «Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks». En: *Computerized Medical Imaging and Graphics* 74 (mar. de 2019). DOI: 10.1016/j.compmedimag.2019.02.003 (vid. pág. 77).

KIRILLOV, Alexander, *et al.* «Segment Anything». En: *arXiv preprint arXiv:2304.02643* (2023) (vid. pág. 23).

LACOUT, Alexis, *et al.* «Pancreatic involvement in hereditary hemorrhagic telangiectasia: assessment with multidetector helical CT». En: *Radiology* 254.2 (2010), págs. 479-484 (vid. pág. 17).

LI, B., *et al.* «Detection of pulmonary nodules in CT images based on fuzzy integrated active contour model and hybrid parametric mixture model». En: *Computational and Mathematical Methods in Medicine* 2013 (2013) (vid. pág. 16).

LI, Gang, *et al.* «Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation». En: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 2. 2022, págs. 1306-1313 (vid. pág. 51).

LI, Zhihui, *et al.* «When object Detection meets knowledge Distillation: A survey». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.8 (mar. de 2023), págs. 10555-10579 (vid. pág. 50).

LIAO, Miao, *et al.* «Pulmonary Nodule Detection from 3D CT Image with a Two-Stage Network». En: *International Journal of Intelligent Systems* 2023 (dic. de 2023), págs. 1-14. DOI: 10.1155/2023/3028869 (vid. pág. 29).

LIU, Shilong, *et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. Mar. de 2023 (vid. págs. 23, 24, 40).

LIU, Wenhu, *et al.* «CSEA-Net: A channel–spatial enhanced attention network for lung tumor segmentation on CT images». En: *iScience* 28.3 (2025) (vid. pág. 31).

LIU, Ze, *et al.* «Swin Transformer: Hierarchical Vision Transformer using Shifted Windows». En: *arXiv (Cornell University)* (ene. de 2021). DOI: 10.48550/arxiv.2103.14030 (vid. pág. 42).

MA, Ling, *et al.* «TiCNet: transformer in convolutional neural network for pulmonary nodule detection on CT images». En: *Journal of Imaging Informatics in Medicine* 37.1 (2024), págs. 196-208 (vid. pág. 31).

MANICKAVASAGAM, R.; SELVAN, S. y SELVAN, Mary. «CAD system for lung nodule detection using deep learning with CNN». En: *Medical & Biological Engineering & Computing* 60.1 (nov. de 2021), págs. 221-228. DOI: 10.1007/s11517-021-02462-3 (vid. pág. 29).

MARTIN, Maria D., *et al.* «Lung-RADS: Pushing the limits». En: *Radiographics* 37.7 (oct. de 2017), págs. 1975-1993. DOI: 10.1148/rg.2017170051 (vid. pág. 15).

MCWILLIAMS, Annette, *et al.* «Probability of cancer in pulmonary nodules detected on first screening CT». En: *New England journal of medicine* 369.10 (2013), págs. 910-919 (vid. pág. 17).

MEI, J., *et al.* «SANet: A Slice-Aware Network for Pulmonary Nodule Detection». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.8 (2021), págs. 4374-4387. DOI: 10.1109/TPAMI.2021.3058744 (vid. pág. 14).

METS, O. M., *et al.* «Incidental perifissural nodules on routine chest computed tomography: lung cancer or not?» En: *European Radiology* 28 (2018), págs. 1095-1101 (vid. pág. 16).

MKINDU, H.; WU, L. y ZHAO, Y. «Lung nodule detection in chest CT images based on vision transformer network with Bayesian optimization». En: *Biomedical Signal Processing and Control* 85 (2023), pág. 104866. DOI: 10.1016/j.bspc.2023.104866 (vid. págs. 13, 30).

MONKAM, Patrice, *et al.* «Detection and Classification of Pulmonary Nodules Using Convolutional Neural Networks: A Survey». En: *IEEE Access* 7 (2019), págs. 78075-78091. DOI: 10.1109/ACCESS.2019.2920980 (vid. pág. 16).

NGUYEN, Tan-Cong, *et al.* «MANet: Multi-branch attention auxiliary learning for lung nodule detection and segmentation». En: *Computer Methods and Programs in Biomedicine* 241 (ago. de 2023), pág. 107748. DOI: 10.1016/j.cmpb.2023.107748 (vid. pág. 29).

OPENAI; ADLER, Steven; AGARWAL, Sandhini, *et al. GPT-4 Technical Report*. Inf. téc. OpenAI, 2024 (vid. pág. 23).

OPULENCIA, Pia, *et al.* «Mapping LIDC, RadLex™, and Lung Nodule Image Features». En: *Journal of Digital Imaging* 24.2 (mar. de 2010), págs. 256-270. DOI: 10.1007/s10278-010-9285-6 (vid. pág. 55).

PAI, Suraj, *et al.* «Foundation model for cancer imaging biomarkers». En: *Nature Machine Intelligence* 6.3 (mar. de 2024), págs. 354-367. DOI: 10.1038/s42256-024-00807-9 (vid. pág. 32).

RUBIN, Geoffrey D. «Lung Nodule and Cancer Detection in Computed Tomography Screening». En: *Journal of Thoracic Imaging* 30.2 (2015), págs. 130-138 (vid. págs. 12, 15).

SETIO, Arnaud Arindra Adiyoso, *et al.* «Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge». En: *Medical Image Analysis* 42 (jul. de 2017). DOI: 10.1016/j.media.2017.06.015 (vid. pág. 77).

SIM, Y., *et al.* «Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs». En: *Radiology* 294.1 (2020), págs. 199-209. DOI: 10.1148/radiol.2019191193 (vid. pág. 13).

TAN, Mingxing y LE, Quoc V. «EfficientNet: Rethinking model scaling for convolutional neural networks». En: *arXiv (Cornell University)* (ene. de 2019). DOI: 10.48550/arxiv.1905.11946 (vid. pág. 52).

TANG, Hao; KIM, Daniel R. y XIE, Xiaohui. «Automated pulmonary nodule detection using 3D deep convolutional neural networks». En: *arXiv (Cornell University)* (ene. de 2019). DOI: 10.48550/arxiv.1903.09876 (vid. pág. 77).

TEAM, The National Lung Screening Trial Research. «Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening». En: *New England Journal of Medicine* 365.5 (2011), págs. 395-409. DOI: 10.1056/NEJMoa1102873 (vid. pág. 55).

VAN RIEL, Sarah J, *et al.* «Observer variability for classification of pulmonary nodules on low-dose CT images and its effect on nodule management». En: *Radiology* 277.3 (2015), págs. 863-871 (vid. pág. 17).

VASWANI, Ashish, *et al.* «Attention is all you need». En: *Advances in neural information processing systems* 30 (2017) (vid. pág. 19).

W. H. ORGANIZATION. *Global cancer burden growing, amidst mounting need for services*. News release. World Health Organization. Feb. de 2024. URL: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services (vid. págs. 12, 15, 33).

WU, Xiaosheng, *et al.* «YOLO-MSRF for lung nodule detection». En: *Biomedical Signal Processing and Control* 94 (abr. de 2024), pág. 106318. DOI: 10.1016/j.bspc.2024.106318 (vid. pág. 30).

XIA, Zhuofan, *et al.* «Vision Transformer with Deformable Attention». En: *arXiv (Cornell University)* (ene. de 2022). DOI: 10.48550/arxiv.2201.00520 (vid. pág. 44).

XU, Jing , *et al.* «An improved faster R-CNN algorithm for assisted detection of lung nodules». En: *Computers in Biology and Medicine* 153 (dic. de 2022), pág. 106470. DOI: 10.1016/j.compbiomed.2022.106470 (vid. pág. 29).

ZHANG, Shifeng, *et al.* «Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection». En: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, págs. 9759-9768 (vid. pág. 21).

ZHAO, Yian, *et al.* «Detrs beat yolos on real-time object detection». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 16965-16974 (vid. págs. 21, 37).

ZHENG, Zhaohui, *et al.* «Localization distillation for dense object detection». En: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, págs. 9407-9416 (vid. págs. 49, 50).