LOCALIZACIÓN DE NÓDULOS PULMONARES UTILIZANDO REPRESENTACIONES PROFUNDAS CONTEXTUALES

DAVID SANTIAGO MORANTES DUARTE
CARLOS ANDRÉS GUTIÉRREZ BENAVIDES

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

LOCALIZACIÓN DE NÓDULOS PULMONATES UTILIZANDO REPRESENTACIONES PROFUNDAS CONTEXTUALES

DAVID SANTIAGO MORANTES DUARTE
CARLOS ANDRÉS GUTIÉRREZ BENAVIDES

Tesis presentada en cumplimiento de los requisitos para optar por el título de:
Ingeniero de Sistemas

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

Codirector:
Luis Carlos Guayacán Chaparro
Magíster en Ingeniería de Sistemas e Informática

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

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a mi director Fabio Martínez, por todo el esfuerzo y dedicación invertidos en este trabajo de investigación. Su paciencia, sabiduría y esa particular gracia que lo caracteriza hicieron que este proceso fuera mucho más llevadero y enriquecedor, dejándome grandes enseñanzas que trascienden el ámbito académico. A mi codirector Luis Guayacán, le agradezco no solo por su guía profesional, sino también por convertirse en un mentor y amigo durante mi recorrido universitario, enseñándome una cantidad excesiva de conceptos a punta de memes. Su apoyo constante y su orientación me ayudaron a encontrar mi camino profesional.

A mi compañero de tesis Carlos Gutiérrez que me permitió observar la importancia del trabajo en equipo así existieran tantos problemas de por medio.

A mi madre Gloria Duarte y mi padre Jesús Morantes que me han apoyado siempre sin importar las adversidades, quienes fueron los responsables de que pudiera cumplir mis sueños a lo largo de tantos semestres y que me han brindado los valores y experiencias que han hecho de mí lo que soy hoy en día. A mi novia María Montt que estuvo presente todo el tiempo, celebrando cada pequeño logro que tenia e incentivándome a conseguir más.

Agradezco también a mis compañeros del grupo de investigación BIVL^{2}ab con los cuales el *transfer learning* fue de vital importancia a la hora de terminar el trabajo de grado.

Finalmente a mis amigos, especialmente a Juan Bermudez y Josué de la Rosa, con los cuales pude compartir muchas experiencias y logros.

- David Santiago Morantes

AGRADECIMIENTOS

A mi director, el profesor Fabio Martínez, Ph.D., expreso mi más profundo agradecimiento por haberme brindado la oportunidad de formar parte del grupo de investigación. Su tiempo, dedicación, guía y paciencia han sido fundamentales en mi formación. Valoro especialmente su comprensión respecto a las particularidades de mis viajes al llano, lo que refleja su empatía y compromiso con mi desarrollo. De igual manera, extiendo mi gratitud a mi codirector, Luis Guayacán, Ph.D.(c), a quien no solo considero un mentor, sino también un amigo. Las reuniones que compartimos, en las que se combinaban temas científicos, discusiones serias y conversaciones con diversas ocurrencias, fueron un espacio enriquecedor que fortaleció el sentido de comunidad dentro del Team detection. Estas experiencias hicieron que mi camino en la ciencia y la investigación fuera no solo más llevadero, sino también profundamente significativo.

A mi compañero de tesis David Morantes por compartir todo este proceso conmigo, tenerme bastante paciencia, y a la vez motivándome a cumplir nuestro objetivo y lograr graduarnos.

A mi madre Luisa Benavides, mi hermano David Gutiérrez, mi novia María Alejandra Cisneros y su familia, quienes han sido mi ancla y motor durante todo este proceso. Su amor incondicional y su fe en mis capacidades han sido mi mayor fortaleza, incluso en los momentos más difíciles.

Agradezco también a mis compañeros del grupo de investigación BIVL^{2}ab por su respaldo y colaboración. Cada experiencia compartida, cada reto superado en equipo, ha enriquecido enormemente este camino. A mis amigos, especialmente a Daniel Cañate, Jaider Mare, Nixon Vargas y Jaime Granados, gracias por su compañerismo y por celebrar conmigo cada paso, cada pequeño logro. Su presencia ha hecho que este camino sea mucho más especial.

Finalmente, agradezco a la Coca Cola.

- Carlos Gutiérrez

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

ABREVIATURAS

Lista unificada de abreviaciones y siglas empleadas en el documento.

Abreviación	**Significado**
AGG	*Aggregation function* (función de agregación).
BERT	Bidirectional Encoder Representations from Transformers.
CA	Cross-Attention.
CLIP	Contrastive Language–Image Pre-training.
CP	Cáncer de pulmón.
DALL·E	(Modelo de generación de imágenes de OpenAI).
DA	Deformable Attention.
DINO	DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection (modelo de auto-atención).
FPN	Feature Pyramid Network.
FP	Falsos positivos.
GCN	Graph Convolutional Network.
GAT	Graph Attention Network.
GNN	Graph Neural Network.
GraphSAGE	SAmple and AggreGatE for large-scale graphs.
MF	Modelo fundacional.
MLM	Masked Language Modeling (pretexto en BERT).
NMS	Non-Maximum Suppression.
NP	Nódulo pulmonar.
NSP	Next Sentence Prediction (pretexto en BERT).

(Continúa en la siguiente página)

(Continuación) Lista de abreviaciones y siglas.

Abreviación	**Significado**
RCNN (R-CNN)	Region-based Convolutional Neural Network.
ResNet	Residual Network.
RFP	Reductor de falsos positivos.
ROI	Region of Interest (región de interés).
RPN	Region Proposal Network.
SAM	Segment Anything Model.
SA	Self-Attention.
SSD	Single-Shot Detector.
TC	Tomografía computarizada.
YOLO	You Only Look Once (modelo de detección de una etapa).

GLOSARIO

Tabla de notación y símbolos matemáticos, ordenados por (1) letras minúsculas, (2) letras mayúsculas, (3) letras griegas, (4) otros símbolos.

Símbolo	Descripción
1. Letras minúsculas	
a_i	Representación escalar/vector minúscula con subíndice i.
c_i	Variable (minúscula c) con subíndice i.
d	Dimensión escalar (p.ej. embedding size).
e^l	Conjunto de aristas del grafo para la capa (o escala) l.
$f(\cdot)$	Función de activación o transformación (ej. ReLU).
h^l	Matriz de representaciones nodales globales del grafo para la capa (o escala) l .
h_i, h_i', h_i''	Representaciones (descriptor) del nodo i.
i, j, k	Índices en minúscula (nodos, vecinos, iteraciones).
m_i	Mensaje propagado desde el nodo i.
n	Índice para candidatos, parches, etc.
s	Número de escalas o variable escalar minúscula.
e^l	Conjunto de nodos del grafo para la capa (o escala) l.
2. Letras mayúsculas	
A	Matriz de adyacencia de un grafo.
C^l	Número de canales (o características) en la capa l.
$G^l = (v^l, e^l)$	Grafo para la capa (o escala) l.

(Continúa en la siguiente página)

(Continuación) Tabla de notación y símbolos.

Símbolo	Descripción
H, W	Altura y anchura.
H', W'	Altura y anchura reescaladas (p.ej. en RFP).
K	Matriz de *keys* (clave) en el mecanismo de atención.
M	Matriz de pesos (transformación lineal).
N	Número total de nodos en un grafo (mayúscula N).
Q	Matriz de *queries* (consulta) en la atención.
V	Matriz de *values* en la atención.
W	Matriz de pesos.
W^T	Matriz W transpuesta.
Z^l, Z^l_{fusion}	Representación de *pooling global* y fusión multiescala.
3. Letras griegas	
α_{ij}, α_k	Coeficiente de atención (entre nodos i, j o escalas k).
ρ	Factor (porcentaje) de dilatación.
$\sigma(\cdot)$	Función de activación (puede ser ReLU, etc.).
Υ	Red de *offset* en atención deformable (mayúscula ípsilon).
$\mathbf{\Phi}^i_{<1,2>}$	Bloques consecutivos (1,2) del Swin Transformer.
$\mathbf{\Psi}^i$	Bloques de atención en la rama textual.
4. Otros símbolos y macros	
a, b	Parámetros auxiliares en $MSA = 4hwC^2 + \{\dots\}$.
$\mathbf{a} \in \mathbb{R}^{2d}$	Vector entrenable para la atención en GAT (en negrita).

(Continúa en la siguiente página)

(Continuación) Tabla de notación y símbolos.

Símbolo	**Descripción**
$\mathbf{C_n}, \mathbf{C'_n}$	Cajas delimitadoras para el n-ésimo candidato y su versión dilatada.
$\mathbf{C_n} = \mathbf{I}(x_n, y_n, w_n, h_n)$	Definición de bounding box en coordenadas (x_n, y_n, w_n, h_n).
$\mathbf{F_I^L}, \mathbf{F_T^L}$	Representaciones (visual, texto) en la capa L.
$\mathbf{F''_I}, \mathbf{F''_T}$	Representaciones refinadas tras módulos de atención.
$\mathbf{I} \in \mathbb{R}^{H \times W \times C}$	Imagen (o volumen) de entrada (dimensiones H, W, C).
$\hat{\mathbf{F}}''_\mathbf{I}$	Subconjunto (queries) de $\mathbf{F''_I}$.
$\mathbf{N_i}$	Conjunto de queries seleccionados (ej. top 900).
$\mathbf{N'_i}, \mathbf{N''_i}$	Representaciones intermedias y finales de queries.
$\mathbf{p} = \{p_1, \ldots, p_n\}$	Parches (tokens) divididos de la imagen.
$\mathbf{T}$	Secuencia de texto de entrada (ej. "A lung nodule in the image").
$\mathbf{V_t}, \mathbf{V_p}, \mathbf{V_s}$	Vectores de embedding (token, posición, segmento).
d_k	Dimensión de K y Q en la atención (para $\sqrt{d_k}$).
x_n, y_n, w_n, h_n	Coordenadas y dimensiones de la caja delimitadora.
AGG$(\cdot)$	Función de agregación (media, suma, etc.).
$\sqrt{d}$	Factor de normalización en la atención (producto punto).
concat$(\cdot)$	Concatenación de múltiples cabezas de atención.
FFN	*Feed-Forward Network* para predecir cajas.
MSA	*Multi-Head Self-Attention* (macro).
softmax	Función de normalización en el mecanismo de atención.
Z^m	Resultado de la m-ésima cabeza de atención.
SA, CA, DA	*Self-Attention*, *Cross-Attention*, *Deformable-Attention*.

RESUMEN

TÍTULO: Localización de nódulos pulmonares utilizando representaciones profundas contextuales [*]

AUTORES: David Santiago Morantes Duarte, Carlos Andrés Gutiérrez Benavides. [**]

PALABRAS CLAVE: Cáncer de pulmón, tomografía computarizada, aprendizaje profundo, generalización, modelos fundacionales, representaciones de grafos.

DESCRIPCIÓN: El cáncer de pulmón (CP) es la principal causa de mortalidad por cáncer en el mundo [1]. En 2020, se reportaron 2,21 millones de casos nuevos y 1,8 millones de muertes por esta enfermedad [1]. El diagnóstico temprano del CP involucra la identificación oportuna de nódulos pulmonares (NP), los cuales son pequeñas lesiones sospechosas de malignidad, detectadas usualmente mediante estudios de tomografía computarizada (TC). Sin embargo, este proceso requiere de la interpretación meramente observacional de las imágenes por parte de los radiólogos, lo que resulta en un proceso subjetivo [2]. Además, los NP representan entre el 0.03% y 0.3% del tamaño total de una imagen de TC [3], y comparten similitudes morfológicas con estructuras anatómicas como los vasos sanguíneos, lo que puede llevar a que sean pasados por alto o a la detección de falsos positivos (FP) [4]. Como soporte para esta tarea, se han desarrollado métodos

---

[*] Trabajo de investigación

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez, PhD en ingeniería de sistemas y computación, análisis de imágenes y análisis de vídeo. Codirector: Luis Carlos Guayacán, Ph.D(c)

[1] Jacques FERLAY et al. "Cancer statistics for the year 2020: An overview". In: *International journal of cancer* 149.4 (2021), pp. 778–789.

[2] Annemilia DEL CIELLO et al. "Missed lung cancer: when, where, and why?" In: *Diagnostic and interventional radiology* 23.2 (2017), p. 118.

[3] Geoffrey D RUBIN. "Lung nodule and cancer detection in computed tomography screening". In: *Journal of thoracic imaging* 30.2 (2015), pp. 130–138.

[4] Narjust DUMA; Rafael SANTANA-DAVILA, and Julian R MOLINA. "Non–small cell lung cancer: epidemiology, screening, diagnosis, and treatment". In: *Mayo Clinic Proceedings*. Vol. 94. 8. Elsevier. 2019, pp. 1623–1640.

computacionales para localizar NP aprendiendo y extrayendo patrones y características texturales [5] [6]. Estos métodos, sin embargo, tienen un sesgo de caracterización local (convoluciones), perdiendo información del contexto en el TC[7] [8]. Este estudio explora la detección de NP en imágenes de TC mediante la combinación de un modelo fundacional (MF) y redes de grafos. Inicialmente, se utiliza Grounding-DINO para generar representaciones visuales globales, capturando el contexto semántico y anatómico de las imágenes. Estas representaciones guían la predicción inicial de NP. Posteriormente, las redes de grafos refinan las predicciones al capturar relaciones espaciales y contextuales entre los NP y las estructuras anatómicas circundantes, lo que contribuye a la reducción de FP. Este enfoque combina la capacidad del MF para identificar características relevantes con la fortaleza de las redes de grafos para modelar el contexto, mejorando la detección de NP.

---

[5] Patrice MONKAM et al. "Detection and classification of pulmonary nodules using convolutional neural networks: a survey". In: *Ieee Access* 7 (2019), pp. 78075–78091.

[6] Jose GEORGE; Shibon SKARIA; VV VARUN, et al. "Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans". In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. SPIE. 2018, pp. 347–355.

[7] Gaël VAROQUAUX and Veronika CHEPLYGINA. "Machine learning for medical imaging: methodological failures and recommendations for the future". In: *NPJ digital medicine* 5.1 (2022), p. 48.

[8] Jie MEI et al. "SANet: A slice-aware network for pulmonary nodule detection". In: *IEEE transactions on pattern analysis and machine intelligence* 44.8 (2021), pp. 4374–4387.

ABSTRACT

TITLE: Localization of pulmonary nodules using deep contextual representations. *

AUTHORS: David Santiago Morantes Duarte, Carlos Andrés Gutiérrez Benavides. **

KEYWORDS: Lung Cancer, Computed Tomografy, Deep Learning, Generalization, Foundational Models, Graph Representations.

DESCRIPTION: Lung cancer (LC) is the leading cause of cancer-related mortality worldwide [1]. In 2020, 2.21 million new cases and 1.8 million deaths were reported due to this disease [1]. Early diagnosis of LC involves the timely identification of lung nodules (LN), which are small lesions suspected of malignancy, usually detected through computed tomography (CT) scans. However, this process relies on the purely observational interpretation of images by radiologists, making it subjective [2]. Moreover, LNs account for only 0.03% to 0.3% of the total size of a CT image [3] and share morphological similarities with anatomical structures such as blood vessels, which can lead to LNs being overlooked or to the detection of false positives (FP) [4]. To assist in this task, computational methods have been developed to localize LNs by learning and extracting patterns and textural features from LNs [56]. However, these methods tend to focus on local characterization (convolutions), losing context information in CT images [78]. This study explores the detection of LNs in CT images by combining a foundational model (MF) and graph networks. Initially, Grounding-DINO is used to generate global visual representations, capturing the semantic and anatomical context of the images. These representations guide the initial LN predictions. Subsequently, graph networks refine the predictions by capturing spatial and contextual relationships between the LNs and surrounding anatomical structures, contributing to FP reduction. This approach leverages the MF's ability to identify relevant features and the graph networks' capacity to model the context, improving LN detection.

---

* Research work

** Faculty of Physics-Mechanics Engineering. School of Systems Engineering and Informatics. Advisor: Fabio Martínez Carrillo, PhD. Computer and systems engineering, medical image analysis and video analysis. Co-advisor: Luis Carlos Guayacan, Ph.D(c)

INTRODUCCIÓN

El cáncer de pulmón (CP) es la principal causa de mortalidad por cáncer en el mundo, reportándose 2,21 millones de casos nuevos y 1,8 millones de muertes por esta enfermedad en el año 2020 [1]. El diagnóstico del CP se centra en la detección oportuna de nódulos pulmonares (NP), lesiones anormales de tejido que pueden llegar a ser malignas y que comúnmente se observan e identifican a través de estudios de tomografía computarizada (TC). La detección oportuna de los NP es fundamental para prevenir la propagación del cáncer y poder brindarle al paciente un tratamiento eficaz, lo que impacta directamente en su expectativa de vida. Sin embargo, los NP son masas muy pequeñas (de 3mm a 30mm de diámetro), por lo que su detección, particularmente en estadios tempranos, resulta desafiante [2].

De hecho, estudios en la literatura han reportado que hasta un 25% de los NP no son detectados por los radiólogos [4]. Lo anterior puede deberse entre otras cosas, a la representación espacial mínima en estudios de TC (0.03% y 0.3% del tamaño de una imagen) [3], así como a las similitudes morfológicas y estructurales que comparten los NP con otras partes del pulmón. Adicionalmente, la detección de NP depende en gran medida de la interpretación observacional de los estudios de TC por parte de los radiólogos, resultando en un proceso altamente subjetivo [23]. Para abordar estos problemas se han desarrollado diversos métodos computacionales para soportar la tarea de detección de NP [1], por ejemplo, las representaciones convolucionales [2] han evidenciado capacidad de detección de NP pero con un sesgo de caracterización local, lo que puede llevar a la pérdida de información del contexto en la TC. Por otra parte, los métodos basados enauto-atención han mostrado ser más robustos al tener

---

[1] Yu GU et al. "A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning". In: *Computers in biology and medicine* 137 (2021), p. 104806.

[2] Yongsik SIM et al. "Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs". In: *Radiology* 294.1 (2020), pp. 199–209.

en cuenta relaciones de largo alcance entre las características de la imagen, lo que les permite capturar mejor el contexto y mejorar la precisión de la detección de NP [3] [4]. Sin embargo, estos métodos suelen requerir de una gran cantidad de datos para su entrenamiento y pueden estar sobreajustados a conjuntos de datos específicos [8].

Este trabajo introduce una arquitectura fundacional que captura el contexto de las observaciones de TC, considerando relaciones anatómicas con significado semántico y acoplando una representación geométrica para apoyar la reducción de falsos positivos (FP) durante la predicción. Inicialmente, se adaptó el modelo fundacional (MF) Grounding-DINO, que constituye una representación visual desde múltiples modelos de auto-atención guiadas por un *prompt* [5]. Posteriormente, se exploran arquitecturas basadas en redes de grafos, las cuales se emplean como una estrategia para reducir FP en las predicciones del MF, aprovechando el contexto espacial y anatómico de los NP observados. Este esquema permite detectar NP capturando relaciones de largo alcance entre diferentes partes de la imagen, que ya cuenta con un contexto espacial adicional. Además, el uso de un MF facilita la transferencia de aprendizaje al aprovechar la gran cantidad de datos con la que ha sido preentrenado, mientras que el reductor de falsos positivos (RFP) basado en grafos permite discriminar entre nódulos y otras estructuras similares al capturar relaciones espaciales y morfológicas de las dos clases, lo que proporciona una mayor exactitud y fiabilidad en las predicciones.

---

[3] Alexey DOSOVITSKIY et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[4] Hassan MKINDU; Longwen WU, and Yaqin ZHAO. "Lung nodule detection in chest CT images based on vision transformer network with Bayesian optimization". In: *Biomedical Signal Processing and Control* 85 (2023), p. 104866.

[5] Shilong LIU et al. *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. 2024. arXiv: 2303.05499 [cs.CV].

1. FUNDAMENTOS Y TRABAJOS PREVIOS

1.1. Cáncer de pulmón y los nódulos pulmonares

El cáncer de pulmón (CP) se posiciona como una de las principales causas de muerte por cáncer en el mundo, con más de 2 millones de nuevos casos y 1,8 millones de muertes asociadas en el año 2020[1]. La detección del CP comienza con la revisión del historial médico del paciente, un examen físico y la evaluación de imágenes radiográficas o tomografía computarizada (TC). Estas imágenes pueden revelar la presencia de masas anormales llamadas nódulos pulmonares (NP), las cuales constituyen el principal indicador de cáncer de pulmón [6]. Sin embargo, los NP son pequeños, con diámetros que se encuentran entre 3 y 30 mm, lo que representan menos del 0.3% del tamaño total de la imagen y al rededor del 0.013% de un volumen de TC [3] (ver Figura 1.d). Adicionalmente, los NP presentan una gran variabilidad morfológica asociada tanto a su densidad como a la forma de sus bordes. La densidad textural del tejido del nódulo es representada por el nivel de atenuación en la imagen y se puede clasificar en tres categorías: sólido, vidrio esmerilado y sub-sólido o parcialmente sólido[7].
En cuanto a su morfología, los NP pueden ser redondos, lobulares, especulados o irregulares. Además, según su ubicación y contexto espacial pueden ser clasificados como nódulos yuxtapleurales, yuxtavasculares o aislados[8]. Los nódulos yuxtapleurales (Figura 1.b) son sólidos y se encuentran cerca de la superficie pleural, presentando características benignas y

---

[6] Sean BLANDIN KNIGHT et al. "Progress and prospects of early detection in lung cancer". In: *Open biology* 7.9 (2017), p. 170070.

[7] Maria D. MARTIN et al. "Lung-RADS: Pushing the Limits". In: *RadioGraphics* 37.7 (2017). PMID: 29053407, pp. 1975–1993. DOI: 10.1148/rg.2017170051. eprint: https://doi.org/10.1148/rg.2017170051.

[8] Shingo IWANO et al. "Computer-aided diagnosis: A shape classification of pulmonary nodules imaged by high-resolution CT". in: *Computerized Medical Imaging and Graphics* 29.7 (2005), pp. 565–570. DOI: https://doi.org/10.1016/j.compmedimag.2005.04.009.

con un porcentaje de incidencia aproximado del 21% [9]. Los yuxtavasculares (Figura 1.c) se encuentran adheridos a los vasos sanguíneos, lo que dificulta su detección, tienen una alta probabilidad de ser malignos y presentan una incidencia del 48.7% aproximadamente[10][11]. Por otro lado, los nódulos aislados (Figura 1.a) son aquellos que están rodeados por tejido pulmonar sin otras anomalías y suelen ser principalmente benignos[5].

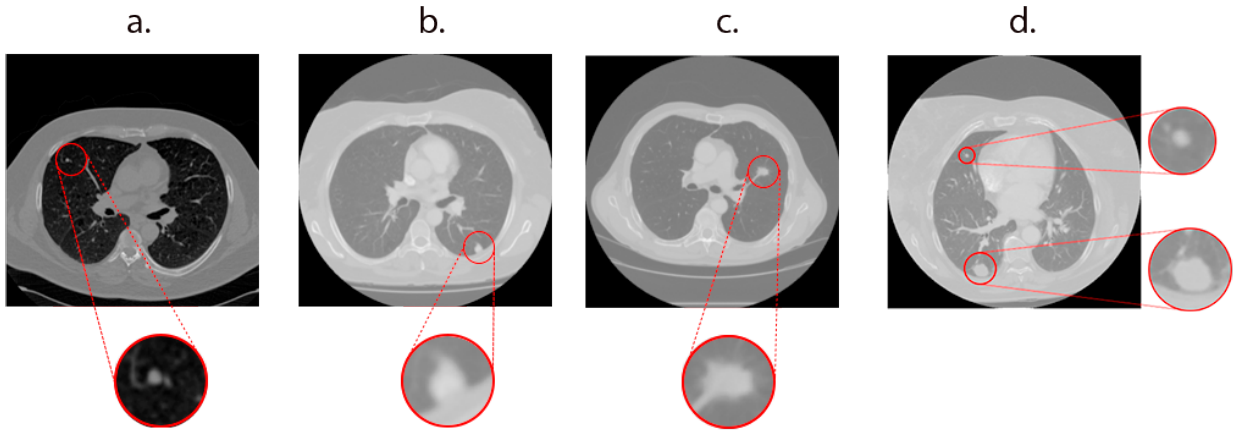

Figura 1. Diferentes tipos de nódulos según su contexto espacial y tamaño a) Nódulo aislado. b) Nódulo yuxtapleural. c) Nódulo yuxtavascular. d) Variabilidad de tamaños entre nódulos pulmonares.

1.2. Estrategias de localización y representaciones contextuales

1.2.1. Arquitecturas para la localización de objetos. En la última década, los modelos de aprendizaje profundo han impulsado de manera significativa los avances en la

---

[9] Onno M METS et al. "Incidental perifissural nodules on routine chest computed tomography: lung cancer or not?" In: *European radiology* 28 (2018), pp. 1095–1101.

[10] Rui HAO; Yan QIANG; Xiaofei 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". In: *Computational and mathematical methods in medicine* 2018 (2018).

[11] Bin LI et al. "Detection of pulmonary nodules in CT images based on fuzzy integrated active contour model and hybrid parametric mixture model". In: *Computational and mathematical methods in medicine* 2013 (2013).

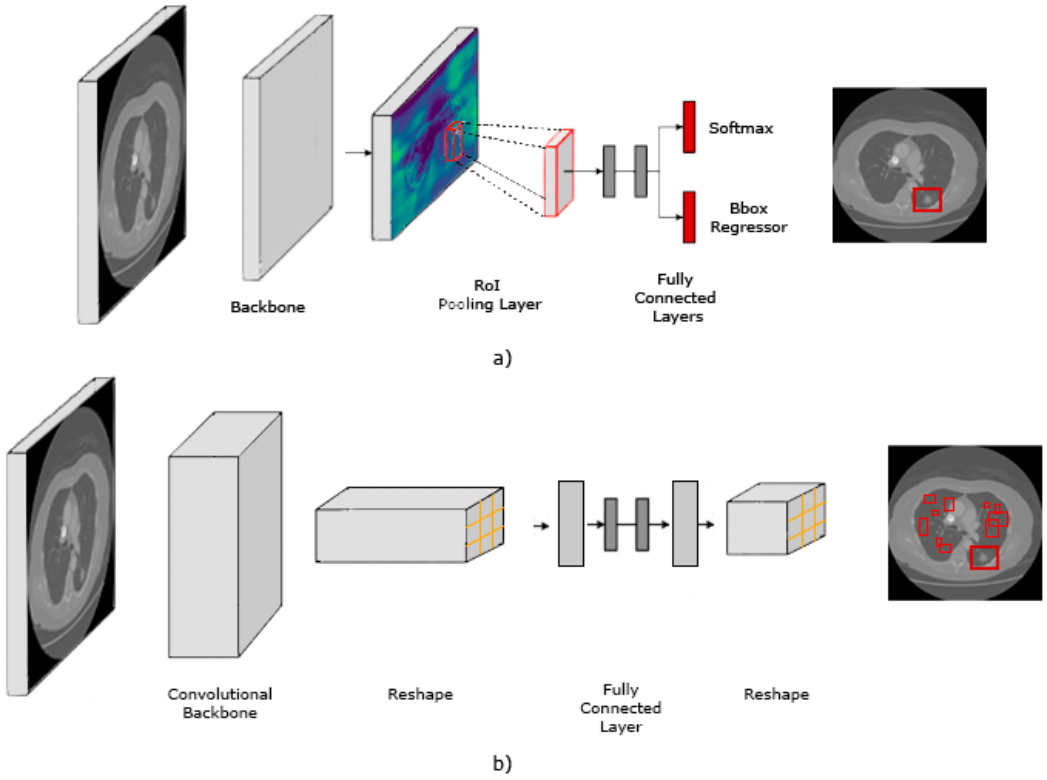

Figura 2. a) Ejemplo de un detector de dos etapas (Fast R-CNN). b) Ejemplo de un detector de una etapa (YOLOv1).

detección de objetos. Estas arquitecturas se dividen generalmente en dos componentes principales. El primero es el módulo de extracción de características, conocido como *backbone*, que genera representaciones profundas a partir de las imágenes de entrada. El segundo es el módulo de detección, cuya tarea es predecir los cuadros delimitadores (o *bounding boxes*, en inglés) que delimitan los objetos de interés. Generalmente, el backbone utiliza redes convolucionales, mientras que los detectores se clasifican en dos categorías principales: detectores de una etapa y detectores de dos etapas. A continuación, se detallan los métodos más relevantes expuestos en la literatura [12].

[12] Syed Sahil Abbas ZAIDI et al. "A survey of modern deep learning based object detection models". In: *Digital Signal Processing* 126 (2022), p. 103514.

Detectores de dos etapas. Los detectores de dos etapas lograron superar las técnicas tradicionales de aprendizaje automático al incorporar redes neuronales convolucionales profundas. Un ejemplo destacado es la Region-based Convolutional Neural Network (R-CNN) [13]. En la primera etapa, se generan propuestas de regiones candidatas, que podrían contener objetos, mediante algoritmos de selección de regiones. En la segunda etapa, las características de estas regiones son extraídas utilizando una CNN preentrenada. Posteriormente, estas características se procesan en dos ramas independientes: una destinada a la clasificación de los objetos y otra para ajustar las coordenadas de las bounding boxes. Al final, se emplea el algoritmo de *non-maximum suppression* (NMS) para eliminar detecciones redundantes [14]. Aunque R-CNN demostró el potencial de las CNNs en la detección de objetos, su implementación resultaba lenta y carecía de la capacidad de entrenamiento end-to-end.

Para solucionar estas limitaciones, se introdujo Fast R-CNN [15], un modelo que optimiza el proceso de inferencia eliminando la necesidad de generar propuestas de regiones de manera previa. Este modelo integra el proceso de "ROI pooling" directamente dentro de la red, lo que facilita la extracción de características de tamaño fijo a partir de las regiones propuestas. Utilizando la arquitectura VGG como backbone, las características extraídas se emplean para clasificar los objetos y ajustar las coordenadas de las cajas delimitadoras.

Posteriormente, Faster R-CNN [16] perfeccionó aún más este proceso al incorporar una Red de Propuesta de Regiones (RPN, por sus siglas en inglés) directamente en la arquitectura,

---

[13] Ross GIRSHICK et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[14] Meiling GONG et al. "A review of non-maximum suppression algorithms for deep learning target detection". In: *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*. Vol. 11763. SPIE. 2021, pp. 821–828.

[15] Ross GIRSHICK. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[16] Shaoqing REN et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149.

eliminando la necesidad de generar propuestas de manera externa. En este modelo unificado, las propuestas de regiones se generan dentro de la misma red, compartiendo características convolucionales con el módulo de detección. Con el uso de ResNet como backbone, Faster R-CNN logra generar representaciones de mayor calidad de las imágenes de entrada. Las propuestas generadas por la RPN se combinan con el ROI pooling para extraer características que luego se procesan para la clasificación y regresión de las bounding boxes. La inclusión de la RPN permitió un notable aumento en la eficiencia y precisión del modelo.

Detectores de una etapa En contraste con los detectores de dos etapas, los modelos de una etapa priorizan la velocidad y la eficiencia computacional al realizar todo el proceso de detección en un solo paso. La arquitectura más destacable de este tipo de detectores es *You Only Look Once* (YOLO) [17], un modelo que se distingue por su capacidad para realizar inferencias en tiempo real. A diferencia de los detectores de dos etapas, YOLO adopta un enfoque unificado: divide la imagen en una cuadrícula y, simultáneamente, predice las cajas delimitadoras, las probabilidades de clase y la presencia de objetos para cada celda de la cuadrícula. Este enfoque permite una detección casi instantánea (alrededor de 22 ms para una imagen) [17]. Sin embargo, YOLO puede enfrentar dificultades cuando se trata de objetos pequeños o escenas con una alta densidad de objetos.

De manera similar, Single-Shot Detector (SSD) [18] es otro modelo destacado que se caracteriza por su rapidez y eficiencia. SSD utiliza mapas de características en diversas escalas para detectar objetos de diferentes tamaños en una única pasada a través de la red, prediciendo simultáneamente las cajas delimitadoras y las probabilidades de clase.

---

[17] Joseph REDMON et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[18] Wei LIU et al. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.

EfficientDet [19] introduce un enfoque de escalamiento compuesto que optimiza tanto la profundidad como el ancho del modelo. Este modelo integra la detección de objetos y la predicción de clases en una única arquitectura, optimizando su backbone mediante una técnica de escalado. EfficientDet ha demostrado ser eficaz para manejar objetos de distintos tamaños y complejidades, destacándose por su precisión y su eficiencia en el uso de parámetros y recursos computacionales, lo que lo convierte en una excelente opción para entornos con limitaciones de recursos o tiempo real.

Por último, RetinaNet [20] emplea una *Feature Pyramid Network* (FPN) como núcleo de su arquitectura, lo que le permite capturar características a diferentes escalas, facilitando la detección de objetos de diversos tamaños en una única etapa. RetinaNet también introduce la función de pérdida focal, diseñada para abordar el desequilibrio de clases en los conjuntos de datos de detección. Al compartir una red común para la clasificación y regresión de bounding boxes, RetinaNet optimiza su eficiencia computacional, destacándose por su equilibrio entre precisión y velocidad, lo que lo convierte en una opción robusta para una amplia variedad de aplicaciones en detección de objetos.

1.2.2. Modelos Fundacionales (MF). como respuesta a las limitaciones en los sistemas tradicionales de aprendizaje profundo, relacionado con los esquemas estrictamente supervisados y con modelamiento local de los datos, las arquitecturas basadas en mecanismos de auto-atención permitieron una gestión más eficiente y paralelizada de las relaciones contextuales.

A partir de estas arquitecturas de atención (denominadas comunmente como *transformers*) surgen modelos fundacionales entrenados con grandes cantidades de datos no etiquetados

---

[19] Mingxing TAN; Ruoming PANG, and Quoc V LE. "Efficientdet: Scalable and efficient object detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10781–10790.

[20] T LIN. "Focal Loss for Dense Object Detection". In: *arXiv preprint arXiv:1708.02002* (2017).

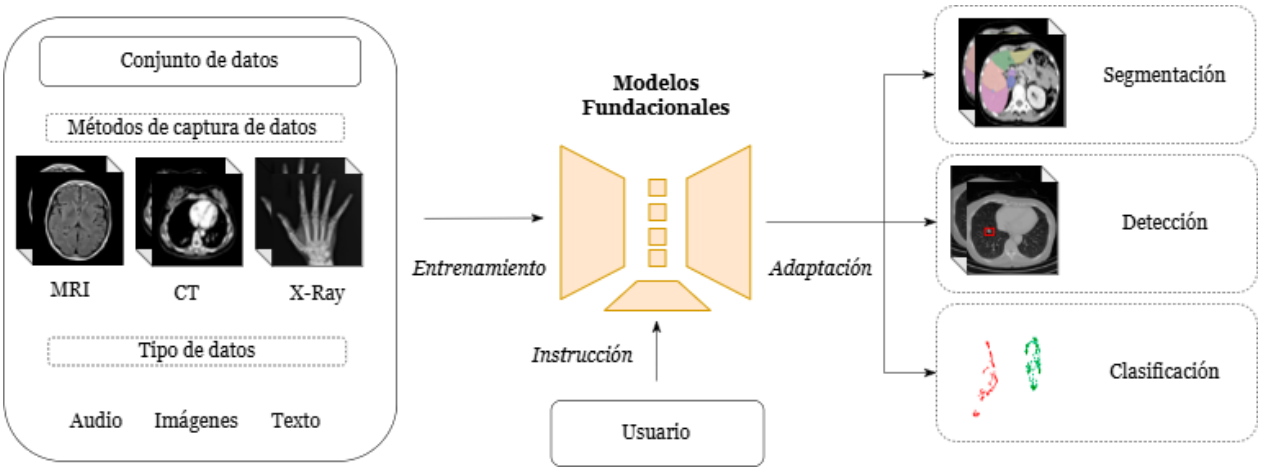

Figura 3. Esquema general del funcionamiento de un MF. Los MF aprenden características de bajo nivel a partir de datos de diversas fuentes y dominios, lo que les permiten realizar diferentes tipos de tareas.

utilizando tareas de pretexto, lo cual permitió la generación de información sin restricción de tareas especificas.

Estos modelos multimodales son capaces de procesar e integrar información proveniente de diversas fuentes, permitiendo realizar tareas como clasificación, detección, segmentación y síntesis de imágenes guiadas por descripciones textuales. Ejemplos de estos modelos incluyen CLIP, YOLO-WORLD, Grounding-DINO, SAM y DALL·E [21][22][23][24]. Con el tiempo, el término *Modelo Fundacional* se adoptó para describir modelos altamente generales y escalables que podían ser preentrenados en grandes cantidades de datos y adaptados a una variedad de tareas específicas. Este enfoque resulta especialmente relevante en ámbitos como la medicina,

---

[21] Alec RADFORD et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[22] Tianheng CHENG et al. *YOLO-World: Real-Time Open-Vocabulary Object Detection*. 2024. arXiv: 2401.17270 [cs.CV].

[23] Alexander KIRILLOV et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV].

[24] Aditya RAMESH et al. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), p. 3.

donde la privacidad de los pacientes y la dificultad para establecer acuerdos de cooperación con entidades médicas complican la adquisición de datos necesarios para entrenar modelos robustos.

Estos modelos presentan características como la escalabilidad, siendo entrenados con volúmenes masivos de datos, lo que les permite capturar patrones y representaciones altamente generales que son aprovechadas para distintas tareas. Por otra parte, también son reconocidos por su robustez a entradas ruidosas y ajustandose a tareas especializadas mediante técnicas como el ajuste fino, el uso de conectores (*bridgers*) y el ajuste basado en instrucciones (*instruction tuning*). El ajuste fino o transferencia de aprendizaje implica adaptar ciertas representaciones del MF utilizando un conjunto de datos pequeño enfocado en el problema específico a resolver. El uso de conectores, que son pequeños bloques entrenables o no entrenables, adapta la información de salida de un MF para que pueda ser utilizada por otras arquitecturas más especializadas, o que preparan la información entrante para brindar contexto al MF. A su vez, el ajuste basado en instrucciones utiliza descripciones textuales para guiar al modelo en el aprendizaje de tareas específicas, reduciendo la necesidad de datos etiquetados adicionales [25]. Es importante destacar que estas estrategias no son mutuamente excluyentes; de hecho, pueden emplearse conjuntamente para mejorar la adaptabilidad del MF [26].

1.2.3. Modelos basados en grafos Las redes neuronales de grafos (*GNN*, por sus siglas en inglés), representan una categoría de modelos de deep learning diseñados para procesar datos estructurados, aprovechando la geometría y contexto de la información, modelados como grafos. Estos grafos son herramientas matemáticas comunes para representar relaciones entre distintas entidades, constituidos por nodos (vértices) asociados a dichas entidades, tales

---

[25] Shengyu ZHANG et al. "Instruction tuning for large language models: A survey". In: *arXiv preprint arXiv:2308.10792* (2023).

[26] Edward J HU et al. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

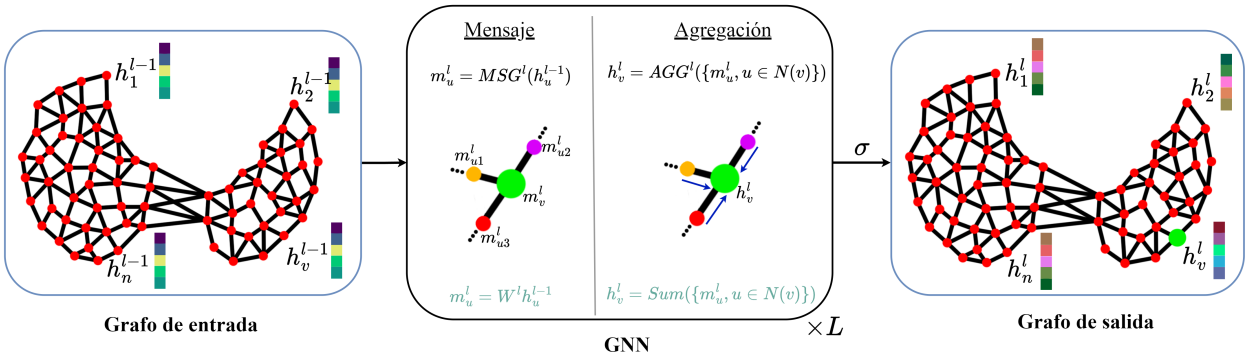

Figura 4. Esquema general del proceso de propagación y agregación de mensajes en una red de grafos. Este proceso permite actualizar iterativamente las representaciones de los nodos al combinar información de sus vecinos en el grafo, capturando tanto la estructura como las características locales y globales.

como los NP y otras estructuras anatómicas, que son interconectados por bordes (aristas) [27] [28].

Las GNN operan en un ciclo de propagación de mensajes. En primer lugar, se lleva a cabo una operación de paso de mensajes donde se calcula una nueva representación para cada nodo v_i multiplicando su representación actual h_i por una matriz de pesos M, expresado como $m_i = M \cdot h_i$. Posteriormente, se agrega la información de los nodos vecinos para obtener una representación agregada a_i, calculada sumando o promediando los mensajes de los vecinos del nodo v_i, expresado como $a_i = \text{AGG}(\{m_j | v_j \in \mathcal{N}(v_i)\})$. Luego, se concatenan estas representaciones agregadas con las representaciones originales de los nodos h_i para capturar tanto la información local como la información contextual, expresado como $c_i = [h_i, a_i]$. Entonces, se aplica una función de activación no lineal σ a la concatenación c_i para obtener la nueva representación del nodo v_i, expresado como $h_i^{\text{new}} = \sigma(c_i)$. Este proceso se repite iterativamente para actualizar las representaciones de los nodos en cada capa de la red.

---

[27] Bharti KHEMANI et al. "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions". In: *Journal of Big Data* 11.1 (2024), p. 18.

[28] Jie ZHOU et al. "Graph neural networks: A review of methods and applications". In: *AI Open* 1 (2020), pp. 57–81. DOI: https://doi.org/10.1016/j.aiopen.2021.01.001.

Una ventaja destacada de las GNN radica en su invariabilidad ante la permutación de nodos, lo que las hace adaptables a grafos con nodos dispuestos en diferentes configuraciones [29]. Esto podría permitir caracterizar eficazmente las estructuras pulmonares, a pesar de su gran variabilidad entre pacientes. Asimismo, estas redes son capaces de manejar grafos de tamaño variable, otorgándoles versatilidad en una amplia gama de contextos y aplicaciones. Con el transcurso del tiempo, las GNN han evolucionado tanto en términos de arquitectura como de rendimiento. Se han desarrollado nuevas capas y técnicas de aprendizaje que posibilitan un modelado más preciso de las relaciones presentes en el grafo. Además, se han explorado enfoques como la atención y la memoria para capturar interacciones de largo alcance y estructuras de red de alta dimensionalidad.

Existen diferentes tipos de GNN, entre las cuales se destacan:

- Graph Convolutional Networks (GCN): Estas redes propagan la información a través de los nodos y bordes del grafo, actualizando iterativamente las representaciones nodales según la información local y global. Esto permite capturar patrones complejos y características cruciales de la estructura del grafo[30].

  la actualización de los nodos se realiza utilizando una operación de convolución en el dominio del grafo. Teniendo una matriz de características de nodos h y una matriz de adyacencia del grafo A. Entonces, la operación de convolución se puede expresar como $h' = f\left(A \cdot h \cdot M^T\right)$ donde h' es la matriz de características de nodos actualizada, M son los pesos de la convolución, y f es una función de activación no lineal.

- Graph Attention Networks (GAT): Las GAT utilizan mecanismos de atención para calcular la importancia relativa de los nodos vecinos durante la propagación de la

---

[29] Benjamin SANCHEZ-LENGELING et al. "A Gentle Introduction to Graph Neural Networks". In: *Distill* (2021). https://distill.pub/2021/gnn-intro. DOI: 10.23915/distill.00033.

[30] Thomas N. KIPF and Max WELLING. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG].

información. Esto permite capturar interacciones de largo alcance y estructuras de alta dimensionalidad en el grafo[31].

Estas redes utilizan atención para calcular la importancia relativa de los nodos vecinos durante la propagación de la información. Para cada nodo i, se calcula un coeficiente de atención a_{ij} para cada vecino j como una función de los atributos de los nodos h_i y h_j Luego, se calcula una suma ponderada de los atributos de los vecinos para actualizar el nodo i . Matemáticamente, esto se expresa como $h_i{}' = \sigma\left(\sum_{j\epsilon N(i)} a_{ij} M h_j\right)$ donde σ es una función de activación y M son los pesos de la red.

- GraphSAGE: Esta arquitectura de red utiliza técnicas de muestreo y agregación para generar representaciones de nodos en grafos grandes y heterogéneos. Utiliza el vecindario de un nodo para generar una representación agregada del mismo, lo que facilita el aprendizaje en grafos de tamaño variable y con estructuras complejas[32]. En GraphSAGE, se utiliza el muestreo y la agregación para generar representaciones de nodos. Para cada nodo i se muestrea un subconjunto de sus vecinos y se realiza una operación de agregación para generar una representación agregada del nodo. Esto se expresa matemáticamente como $h_i{}' = AGG\left(\{h_j, \forall j \epsilon N\left(i\right)\}\right)$ donde AGG es una función de agregación, como la media o la suma, y h_j son los atributos de los vecinos del nodo i.

1.3. Esquemas computacionales para la localización de nódulos

Durante la última década, se han propuesto diversos métodos computacionales para soportar la localización de NP en imágenes de TC y proporcionar herramientas para el diagnóstico oportuno del CP. Estos métodos se pueden agrupar en dos enfoques principales: los que

---

[31] Petar VELIČKOVIĆ et al. *Graph Attention Networks*. 2018. arXiv: 1710.10903 [stat.ML].

[32] Will HAMILTON; Zhitao YING, and Jure LESKOVEC. "Inductive Representation Learning on Large Graphs". In: *Advances in Neural Information Processing Systems*. Ed. by I. GUYON et al. Vol. 30. Curran Associates, Inc., 2017.

emplean etapas independientes para la representación y la localización, y los que implementan arquitecturas completamente integradas.

En cuanto a los métodos que usan etapas independientes de representación y localización, Ying et.al adoptaron una versión modificada del algoritmo Faster R-CNN para la detección de nódulos pulmonares en donde se implementó un método de optimización alternativa para el entrenamiento del modelo. Este enfoque implicaba un proceso iterativo de entrenamiento de la Red de Propuesta de Regiones (RPN) y Fast R-CNN, utilizando las regiones sugeridas por cada etapa para afinar la siguiente [33]. Shen et. al, de igual forma, proponen un sistema de detección semi-supervisado (Semi-CADe) que localiza nódulos pulmonares 3D basados en datos no etiquetados mediante aprendizaje adversario para reducir la escasez de etiquetas [34]. Así mismo, se han propuesto mejoras en la arquitectura Faste R-CNN, utilizando características de múltiples escalas, ajuste adaptativo de la intensidad de entrenamiento y mejoras del campo visual para realzar características globales [35]. Jain et.al presentan un modelo convolucional de dos etapas en donde, en primera instancia, utiliza una U-net para la segmentación delos NP's y la segunda etapa reduce los FP, mejorando la eficiencia y precisión del sistema[36]. Por otra parte, Jerome et.al proponen un método para la detección de nodulos, mejorando inicialmente la imagen de entrada con un filtro mediano adaptativo junto con la técnica de transformada wavelet compleja en tiempo discreto. Posterior a ello, para la segmentación, se

---

[33] Ying SU; Dan LI, and Xiaodong CHEN. "Lung nodule detection based on faster R-CNN framework". In: *Computer Methods and Programs in Biomedicine* 200 (2021), p. 105866.

[34] Zhiqiang SHEN et al. "WS-LungNet: A two-stage weakly-supervised lung cancer detection and diagnosis network". In: *Computers in Biology and Medicine* 154 (2023), p. 106587. DOI: https://doi.org/10.1016/j.compbiomed.2023.106587.

[35] Jing XU et al. "An improved faster R-CNN algorithm for assisted detection of lung nodules". In: *Computers In Biology And Medicine* 153 (2023), p. 106470.

[36] Sweta JAIN; Pruthviraj CHOUDHARI, and Mahesh GOUR. "Pulmonary lung nodule detection from computed tomography images using two-stage convolutional neural network". In: *The Computer Journal* 66.4 (2023), pp. 785–795.

aplica una interpretación topológica y se emplea un clasificador RCNN[37]. Sin embargo, para lograr una sensibilidad competitiva, estos métodos requieren generar una gran cantidad de predicciones, ocasionando en consecuencia una alta tasa de FP.

En cuanto a los métodos de representación y localización de una única etapa. Por ejemplo, el propuesto por Kehong et.al el cual adapta la arquitectura YOLOv5 (you only look once) e introduce tres mejoras significativas al algoritmo original. En primer lugar, utiliza una red de agrupación piramidal espacial basada en el método de agrupación estocástica. En segundo lugar, se aplica una red piramidal bidireccional para la fusión de características a múltiples escalas y mejora la función de pérdida adoptando la función EIoU para optimizar el modelo de entrenamiento[38]. También Xiaosheng et.al se basaron en la metodología de YOLOv7, diseñando una capa de detección de objetos pequeños que se enfoca en extraer características de los NP. Segundo, crean un módulo de campo receptivo a múltiples escalas que extrae características alrededor de los nódulos. Por último, desarrollan un modelo de convolución omni-dimensional que permite ponderar la atención en los datos de entrada[39]. De manera similar, Mammeri et.al utilizan la YOLOv7 debido a que esta proporciona predicciones más rápidas y precisas en comparación con versiones anteriores de YOLO, seguido de un proceso de clasificación multicategoría de los nódulos detectados utilizando el modelo VGG16. Este enfoque les permite identificar nódulos benignos, sospechosos y malignos [40]. Yashar et.al, por su parte, desarrollaron un enfoque jerárquico en donde se utiliza el modelo preentrenado YOLOv5s para detectar los NP. Luego, implementan una etapa de posprocesamiento

---

[37] S ALBERT JEROME et al. "Watershed segmentation with CAFIS and RCNN classification for pulmonary nodule detection". In: *IETE Journal of Research* 69.8 (2023), pp. 5052–5063.

[38] Kehong LIU. "Stbi-yolo: A real-time object detection method for lung nodule recognition". In: *IEEE Access* 10 (2022), pp. 75385–75394.

[39] Xiaosheng WU et al. "YOLO-MSRF for lung nodule detection". In: *Biomedical Signal Processing and Control* 94 (2024), p. 106318.

[40] Selma MAMMERI et al. "Early detection and diagnosis of lung cancer using YOLO v7, and transfer learning". In: *Multimedia Tools and Applications* 83.10 (2024), pp. 30965–30980.

utilizando un clasificador basado en convoluciones 3D, con el fin de lograr una detección precisa[41]. Aunque estos enfoques basados en una o más etapas han demostrado buenos resultados en la detección de NP, presentan limitaciones como la pérdida de información contextual, inherente a la naturaleza local de las convoluciones utilizadas en los clasificadores y detectores tradicionales.

Para superar estas limitaciones, se han desarrollado arquitecturas avanzadas que integran mecanismos de atención y arquitecturas transformers. Por ejemplo, Mkindu et.al propusieron una arquitectura 3D-NodViT basada en ViT con una optimización Bayesiana para la detección de nódulos pulmonares en imágenes de TC, obteniendo un mejor rendimiento con menos recursos que otras arquitecturas profundas[4]. De manera similar, Ramezani et.al presentaron Lung-DETR, un modelo basado en *Deformable Detection Transformer* diseñado para manejar la detección de nódulos pulmonares escasos en datos de TC, formulando el problema como una tarea de detección de anomalías. Este modelo integra proyecciones de intensidad máxima para mejorar la visibilidad de los nódulos, atención deformable para localizar nódulos pequeños y, *focal loss* para priorizar detecciones desafiantes. Lung-DETR demostró ser efectivo en la identificación precisa de nódulos, mostrando alta precisión y sensibilidad incluso en contextos anatómicamente complejos [42]. Este modelo ha sido validado con 9,676 cortes de tomografía computarizada con nódulos mayores de 7mm. Este enfoque puede estar limitado para la detección de nódulos en etapas tempranas, con tamaños relativamente pequeños. Además, en el trabajo no se evidencia como su comportamiento puede ser mapeado en otros escenarios, con otros conjuntos de datos.

---

[41] Yashar AHMADYAR et al. "Hierarchical approach for pulmonary-nodule identification from CT images using YOLO model and a 3D neural network classifier". In: *Radiological physics and technology* 17.1 (2024), pp. 124–134.

[42] Hooman RAMEZANI; Dionne ALEMAN, and Daniel LÉTOURNEAU. "Lung-DETR: Deformable Detection Transformer for Sparse Lung Nodule Anomaly Detection". In: *arXiv preprint arXiv:2409.05200* (2024).

Reductores de falsos positivos (RFP). Uno de los mayores desafíos en la detección de NP es la alta tasa de FP. Estos se presentan cuando el sistema clasifica erróneamente estructuras anatómicas normales del pulmón o irregularidades propias de la imagen como nódulos pulmonares, lo que puede llevar a diagnósticos incorrectos y aumentar innecesariamente los procedimientos médicos [43]. Para abordar este problema, se han implementado diversos métodos, conocidos como reductores de FP, que filtran las detecciones erróneas para mejorar la precisión, procurando no afectar negativamente la sensibilidad[44]. Por ejemplo, El-Regaily et al. desarrollaron un método que combina un clasificador basado en ingeniería de características con una red neuronal convolucional multi-vista. En su enfoque, se extraen características básicas de los nódulos como longitud de ejes, área, volumen y desproporción esférica para definir umbrales que eliminan rápidamente casos obvios de no-nódulos. Su modelo CNN multi-vista, consta de tres ramas separadas con capas convolucionales, de *max-pooling* y capas densas, mejorando significativamente la capacidad del método para diferenciar nódulos reales de hallazgos irrelevantes[45]. También, Zhu et al. propusieron MR-Forest, una versión mejorada del Deep Forest, para la reducción de falsos positivos en tomografías volumétricas. Este método utiliza un mapeo de coordenadas de espacio cartesiano a esférico, mediante indices armónicos esféricos y ventanas de facetas de anillos ordenados, para extraer características clave de ubicación, forma y textura. Las características extraídas se procesan en múltiples niveles de cascadas, integrando salidas ponderadas para optimizar la precisión [46].

---

[43] Juanyun MAI et al. "Mhsnet: Multi-head and spatial attention network with false-positive reduction for lung nodule detection". In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2022, pp. 1108–1114.

[44] Jinglun LIANG et al. "Reducing False-Positives in Lung Nodules Detection Using Balanced Datasets". In: *Frontiers in Public Health* 9 (2021). DOI: 10.3389/fpubh.2021.671070.

[45] Salsabil Amin EL-REGAILY et al. "Multi-view Convolutional Neural Network for lung nodule false positive reduction". In: *Expert systems with applications* 162 (2020), p. 113017.

[46] Hongbo ZHU et al. "MR-forest: a deep decision framework for false positive reduction in pulmonary nodule detection". In: *IEEE Journal of Biomedical and Health Informatics* 24.6 (2019), pp. 1652–1663.

Con el objetivo de capturar información tridimensional, Zheng et al. desarrollaron una estrategia conjunta que integra dos redes CNN 3D, denominadas Archi-2 y Archi-3, las cuales comparten una estructura basada en la VGG-net, pero difieren en la cantidad de capas de convolución y tamaños de kernel para procesar parches cúbicos de distintos tamaños, logrando así una mejor clasificación de nódulos pequeños [47]. Sin embargo, estos métodos se encuentran limitados a la captura de información local por la naturaleza propia de las convoluciones, lo que puede resultar en clasificadores sub-óptimos al ignorar las relaciones de largo alcance. Para solucionar los problemas asociados a las CNN respecto a la captura del contexto global, Sun et al. presentaron una propuesta que integra representaciones convolucionales con modulos de atención, denominados flujos complementarios. Su arquitectura utiliza dos flujos complementarios (zoom-in y zoom-out) para aprender características internas y contextuales de los nódulos. Incorporando un Módulo de Atención de Bloques Convolucionales, para refinar la extracción de características a diferentes escalas, mejorando la reducción de FP [48]. Más recientemente, Gu et al. propusieron una estrategia basada en una CNN 3D que aborda los desafíos de escala variable y distribución irregular de nódulos pulmonares. Utilizan un modelo 3D SENet50 con bloques de compresión y excitación (SE blocks), que recalibran las relaciones entre canales de las características extraídas. Además, implementan un módulo de fusión multiescala guiado por atención cruzada para combinar características a diferentes escalas y un clasificador DisAlign, que mejora el manejo del desequilibrio de clases mediante un enfoque en dos etapas con una pérdida ponderada[49]. Aunque los métodos previamente mencionados han demostrado ser efectivos en la reducción de FP,

---

[47] Sunyi ZHENG et al. "Automatic Pulmonary Nodule Detection in CT Scans Using Convolutional Neural Networks Based on Maximum Intensity Projection". In: *IEEE Transactions on Medical Imaging* 39.3 (2020), pp. 797–805. DOI: 10.1109/TMI.2019.2935553.

[48] Lingma SUN et al. "Attention-embedded complementary-stream CNN for false positive reduction in pulmonary nodule detection". In: *Computers in Biology and Medicine* 133 (2021), p. 104357.

[49] Zhongxuan GU et al. "Cross attention guided multi-scale feature fusion for false-positive reduction in pulmonary nodule detection". In: *Computers in Biology and Medicine* 151 (2022), p. 106302.

aún enfrentan ciertas limitaciones. A pesar de su capacidad para filtrar detecciones redundantes, extraer características relevantes o aplicar mecanismos de atención a los casos más desafiantes, estos modelos pueden ser sensibles a la variabilidad del contexto espacial y a la complejidad de las estructuras pulmonares. Además, suelen requerir grandes cantidades de datos para entrenarse de manera eficiente y pueden experimentar dificultades cuando se enfrentan a conjuntos de datos clínicos escasos [50]. Esto puede resultar en una reducción subóptima de los FP en ciertos escenarios, especialmente cuando la representación contextual es insuficiente o los datos no reflejan adecuadamente la diversidad de casos clínicos [51].

En este contexto, las redes neuronales basadas en grafos, centrados en el modelamiento del contexto global de las imágenes médicas, tienen la capacidad de capturar relaciones espaciales y patrones de datos que los métodos tradicionales pueden pasar por alto [5]. Al integrar representaciones profundas y contextualizadas, estos modelos no solo podrían mejorar la precisión en la detección de NP, sino también lograr una reducción aún más significativa de los FP. Su habilidad para procesar y entender el contexto global de las imágenes los convierte en una opción viable y potencialmente más robusta frente a las limitaciones de las redes tradicionales de detección, especialmente en escenarios clínicos complejos[52].

---

[50] Mohammad A. THANOON et al. "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images". In: *Diagnostics* 13.16 (2023). DOI: 10.3390/diagnostics13162617.

[51] Ioannis MARINAKIS; Konstantinos KARAMPIDIS, and Giorgos PAPADOURAKIS. "Pulmonary Nodule Detection, Segmentation and Classification Using Deep Learning: A Comprehensive Literature Review". In: *BioMedInformatics* 4.3 (2024), pp. 2043–2106.

[52] Lin ZHANG et al. "Graph neural networks for image-guided disease diagnosis: A review". In: *iRADIOLOGY* 1.2 (2023), pp. 151–166.

2. PROBLEMA DE INVESTIGACIÓN

El CP es el tipo de cáncer más prevalente y mortal a nivel mundial [153]. Las TC permiten la localización y caracterización de los NP, principales precursores del CP. Sin embargo, la localización de los NP es una tarea netamente observacional que enfrenta desafios debido a una alta variabilidad de características morfológicas y de contexto espacial que dificultan su detección[2]. Por ejemplo, los NP varían considerablemente en tamaño entre 3 y 30 mm, representando, en un estudio convencional TC, menos del 0.013% [3]. Además existen notables similitudes con estructuras pulmonares, como los vasos sanguíneos. Esta similitud y diferencias de tamaño, sumado con las diversas estructuras torácicas de los pacientes, conducen a que hasta un 25% de los nódulos pasen desapercibidos durante la revisión de los radiólogos [4]. Además, la subjetividad de los radiólogos y el sesgo de búsqueda pueden contribuir a pasar por alto NP durante la interpretación de las TC.

Hoy en dia arquitecturas convolucionales han mostrado ser competitvas en la tarea de localización de NP. Sin embargo, estas estrategias pueden perder información contextual, lo que reduce su capacidad para capturar la complejidad morfológica de los nódulos pulmonares en las TC. En este sentido, se han propuesto métodos basados en atención, los cuales logran capturar relaciones de largo alcance entre las características de la imagen, lo que les permite preservar el contexto y, en consecuencia, mejorar la precisión de la detección de NP. No obstante, estos métodos suelen requerir una mayor cantidad de datos, y sus arquitecturas son susceptibles al sobreajuste. De esta manera, surge la siguiente pregunta de investigación:

---

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

¿Cómo contribuyen las arquitecturas contextuales desde modelos fundacionales y modelos basados en grafos en la localización de nódulos pulmonares?

3. OBJETIVOS

3.1. Objetivo general

Implementar arquitecturas contextuales que aprovechen relaciones estructurales en la tomografía para la localización de nódulos pulmonares.

3.2. Objetivos específicos

- Seleccionar un conjunto de datos de pulmones que contenga secuencias de tomografía computarizada y anotaciones asociadas con la localización de nódulos pulmonares
- Implementar un modelo fundacional para adaptarlo a la localización de nódulos pulmonares aprovechando la información contextual.
- Implementar una representación geométrica para explotar la información contextual de los nódulos pulmonares.
- Validar los modelos contextuales implementados en cuanto a la capacidad de localizar nódulos en secuencias de tomografías computarizadas.

4. MÉTODO PROPUESTO

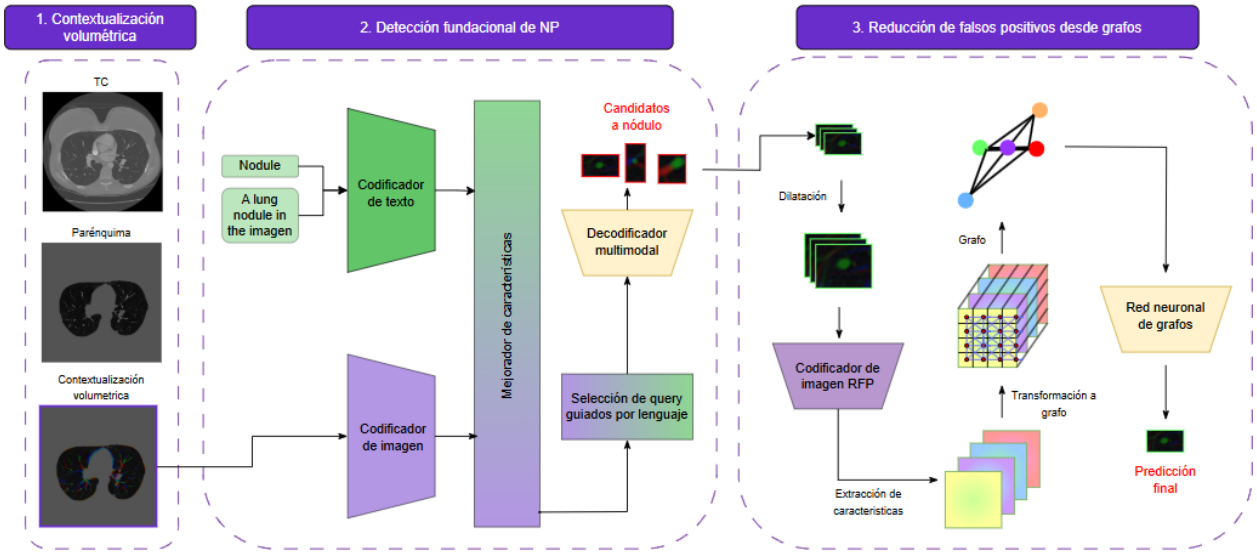

Figura 5. Esquema de la arquitectura propuesta, compuesta de 3 etapas. **(1) Preprocesamiento** En la primera etapa se elimina el parenquima y se le aplica una contextualización volumétrica a las imágenes TC. **(2) Detector de candidatos** en la segunda etapa se hace uso de Grounding-DINO para la predicción de candidatos de NP. **(3) Reductor de falsos positivos** en la tercera etapa, las predicciones anteriores son procesadas por un modelo de grafos con el fin de determinar si realmente corresponden a NP.

Este trabajo propone el uso, ajuste e implementación de una arquitectura fundacional para la detección de NP en imágenes de TC, incluyendo un módulo de reducción de FP. Tanto la representación fundacional como el módulo reductor de FP considera relaciones anatómicas y representaciones geométricas, utilizando modulos de multiple atención y esquemas basados en grafos, respectivamente. La estrategia propuesta primero aplica un preprocesamiento que realza las características contextuales integrando información volumétrica en cada imagen (contextualización volumétrica). Segundo, se ajusta una arquitectura fundacional de tipo Grounding-DINO que preserva el contexto y relaciones estructurales del TC mediante mecanismos de auto-atención y siendo guiado por información de alto nivel textual (*prompt*), que permite la detección inicial de candidatos.

En la tercera etapa, se clasifican los parches de las predicciones obtenidas entre “nódulo” y “no nódulo” buscando patrones locales característicos de la enfermedad. Cómo entrada para este módulo se usan potenciales observaciones, extraidas y ponderadas durante la el procesamiento fundamental para la localización. Posteriormente una represenación basada en grafos incorpora la información contextual de las estructuras adyacentes a cada candidato, lo que permite distinguir con mayor facilidad los nódulos de otras estructuras.

4.1. Contextualización volumétrica en slices TC

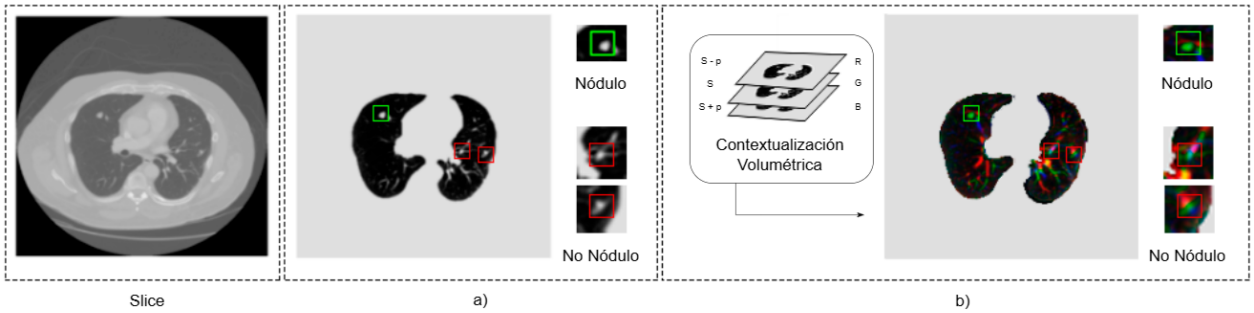

Figura 6. Preprocesamiento de las imágenes TC. a) Se elimina el fondo de los slices. b) Se crea una imagen RGB a partir de 3 slices consecutivos (contextualización volumétrica).

Para el conjunto de datos se realizó una etapa de preprocesamiento compuesta de 2 pasos. Primero, se eliminó el fondo de cada slice de TC utilizando una U-Net preentrenada para segmentar el parénquima pulmonar [54]. Este paso elimina las estructuras externas a los pulmones, como los músculos, los huesos y demás órganos, permitiendo que el modelo se centre únicamente en el contenido del parénquima pulmonar (Figura 6-a). Esta etapa de preprocesamiento ha sido típicamente aplicada en diversas aproximaciones para la localización de NP

[54] Johannes HOFMANNINGER et al. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem”. In: *European Radiology Experimental* 4 (2020), pp. 1–13.

Particularmente la U-Net implementada en este trabajo, fue entrenada con el conjunto de datos R-231, que incluye 108248 imágenes de TC obtenidas con 22 combinaciones diferentes de escáneres, kernels convolucionales y resolución espacial entre *slices* (imágenes 2d que en conjunto representan el volumen). Adicionalmente, sus autores utilizaron un aumento de datos basado en el conjunto *Anatomy3* para incorporar información del tórax en regiones más alejadas de la zona abdominal, con cambios notables en términos morfológicos. El uso combinado de estos conjuntos de datos asegura que el modelo abarque una amplia variabilidad visual, optimizando su capacidad para segmentar de manera precisa el parénquima pulmonar en diferentes contextos clínicos y técnicos[54]. En la figura 6-a se puede observar la segmentación obtenida por esta arquitectura, mostrando entre otras, la reducción de complejidad en el fondo y elementos que no están asociados al parenquima pulmonar. De esta manera, se puede garantizar que la subsiguiente estrategia encargada de la localización de NP estará enfocada en la búsqueda de elementos dentro del parenquima, sin perder compatibilidad por integrar relaciones contextuales complejas, fuera del parénquima.

Al observar un único slice del scan, ciertas estructuras pulmonares, como vasos sanguíneos o bronquios, pueden presentar características visuales muy similares a las de un nódulo (Figura 6-a). Sin embargo, cuando se analiza en un contexto volumétrico, estas estructuras adquieren una morfología diferente, típicamente ramificada, lo que contrasta con la apariencia generalmente esferoidal de los NP. Sin embargo, el tratamiento volumétrico de estos estudios CT, reduce considerablemente el número de datos de entrenamiento, mientras aumenta la

---

[55] Rui XU et al. "Sgda: towards 3d universal pulmonary nodule detection via slice grouped domain attention". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2023).

[56] He MA et al. "Automatic pulmonary ground-glass opacity nodules detection and classification based on 3D neural network". In: *Medical Physics* 49.4 (2022), pp. 2555–2569.

[57] Michael MAYNORD et al. "Semi-supervised training using cooperative labeling of weakly annotated data for nodule detection in chest CT". in: *Medical Physics* 50.7 (2023), pp. 4255–4268.

complejidad en el aprendizaje de relaciones texturales y estructurales del pulmón.
Teniendo en cuenta lo anterior, en este trabajo se integró en cada imagen la información de varios slices consecutivos, facilitándole así al modelo la identificación de los NP respecto a otras estructuras similares. Esta "contextualización volumétrica" se obtiene al hacer una concatenación entre cortes consecutivos, incorporando cortes adyacentes, con cierto paso transversal p (típicamente $p = 1$ para cortes consecutivos), formando una imagen de 3 canales. Esto permite que la imagen obtenida contenga información volumetrica local, sobre slices adyacentes, lo cual resulta valioso para ser procesada directamente por modelos pre-entrenados en imágenes naturales, aprovechando así las características generales de bajo nivel aprendidas en otros dominios. Con este proceso se logra mejorar considerablemente la visibilidad de las variaciones morfológicas, como se observa en la Figura 6-b.

4.2. Detección fundacional de NP

En este trabajo se implementó y adaptó un modelo fundacional, basado en la arquitectura Grounding-DINO [5], para la detección, modelamiento y localización de NP. Este modelo ha sido evaluado exitosamente en diferentes escenarios, siendo entrenado por su carácter fundacional con múltiples conjuntos, permitiendo la detección de objetos en imágenes naturales, guiada por texto. Entre los conjuntos de datos utilizados para el entrenamiento de la arquitectura original, reportada en el estado del arte, se encuentran:

- Dataset COCO. Este conjunto de datos cuenta con 328,000 imágenes y etiquetas asociadas a 2,5 millones de instancias de objetos. Para su uso en la arquitectura Grounding-DINO se usaron la totalidad de imágenes y la información relacionada con detección de objetos, detección de puntos clave y generación de descripciones.

- Dataset O365. Este conjunto de datos cuenta con 600,000 imágenes y etiquetas asociadas a 365 categorías de objetos. Para su uso en la arquitectura Grounding-DINO se usaron la totalidad de imágenes y la información relacionada con 10 millones de cajas

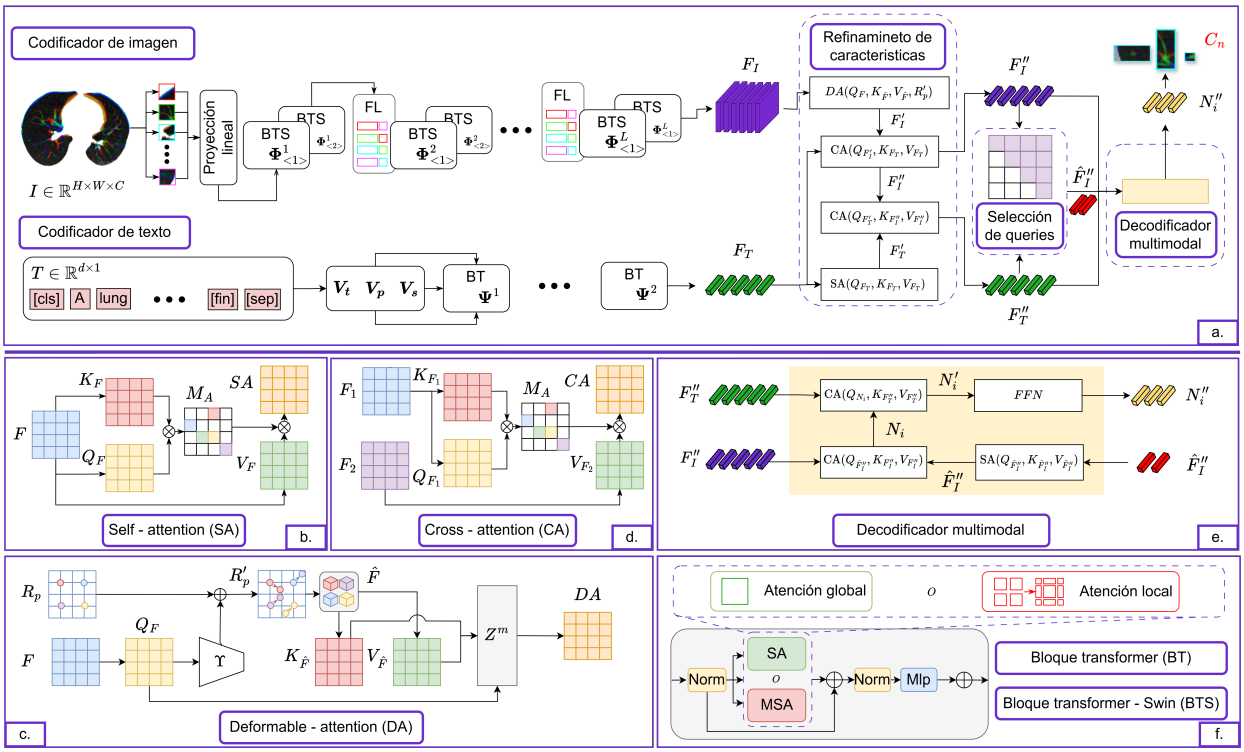

Figura 7. Arquitectura del modelo fundacional Grounding-DINO a), la cual integra características de los codificadores de imagen y texto en un módulo de mejora de características. Este módulo actualiza las representaciones de ambas modalidades, haciendo uso de módulos de auto-atencion b), atención deformable c) y atención cruzada d) facilitando la selección de los *queries* más representativos. Posteriormente, un decodificador multimodal e) procesa estos *queries* refinadas para identificar candidatos a nódulos.

delimitadoras.

- Dataset LVIS. Este conjunto de datos cuenta con 164,000 imágenes con alrededor de 1000 objetos anotados. Este dataset también cuenta con información textual, relacionada con las instancias segmentadas en el dataset repartidos en más de 1000 categorías. Para su uso en la arquitectura Grounding-DINO se usaron la totalidad de imágenes y la información textual relacionada con segmentación de instancias adaptados para la detección.

- Dataset V3Det. Este conjunto de datos cuenta con 13,204 categorías de objetos, repre-

sentadas en un total de 243,000 imágenes. Para su uso en la arquitectura Grounding-DINO se usaron la totalidad de imágenes y la información relacionada con 1,796,818 millones de cajas delimitadoras.

- Dataset GRIT-200K. Este conjunto de 200k imágenes cuenta con pares de imágenes y categorias con anotaciones de ubicación y texto. Para su uso en la arquitectura Grounding-DINO se usaron la totalidad de imágenes y la información relacionada con las anotaciones de ubicación.

- Dataset Flickr30k. Este conjunto de datos cuenta con 31,783 imágenes y etiquetas asociadas a 244,000 cadenas de texto correferenciadas y 276,000 cajas delimitadoras anotadas manualmente. Para su uso en la arquitectura Grounding-DINO se usaron la totalidad de imágenes y la información relacionada con las descripciones textuales y las cajas delimitadoras

Considerando la información anterior, el modelo fundacional base, que esta reportado en la literatura, fue entrenado con una totalidad aproximada de 14,573,818 objetos, observados en 1,566,783 imágenes. Además de esto, el entrenamiento consideró información textual asociada en cada uno de los conjuntos de datos.

En cuanto a la arquitectura, este modelo es inspirado en módulos que indexan múltiples mecanismos de atención (estructura de los *transformers*), permitiendo así retener diversos y complejos patrones desde una gran cantidad de datos. Así, esta arquitectura se compone de diversos módulos diseñados para procesar y relacionar información visual y el contexto que aporta la información textual, en un solo flujo de trabajo. En este trabajo se fijo la entrada textual, considerando que el principal interés es validar y modelar la información principalmente obtenida de las tomografías para la localización de NP.

Particularmente, En la figura 7, cuadro a), se ilustra un esquema general de la arquitectura Grounding-DINO que se adaptó en nuestro contexto, con el objetivo de realizar la detección de NP. Inicialmente se define un codificador multi-escala no local de las observaciones pul-

monares (ver Figura 7- a)). En este caso, se utiliza como entrada una imagen $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, donde $(H \times W)$ son las dimensiones espaciales, mientras que C hace referencia a la contextualización volumétrica, siendo típicamente 3 canales consecutivos. La entrada $\mathbf{I}$ es entonces dividida en n parches (este trabajo uso $n = 16,384$) no superpuestos de tamaño fijo, haciendo una división en grilla de la imagen. Así cada parche tiene una dimensión de 4×4. Cada uno de estos parches se proyecta a una capa lineal para obtener un vector embebido por parche (*token*), como $\mathbf{p} = p_1, p_2, \ldots p_n$. Esta representación de *tokens* es mapeada a una representación de pares de bloques consecutivos *Swin Transformer* (BST), expresando cada par como $\mathbf{\Phi}^i_{<1,2>}$, siendo i la etapa de procesamiento. Entonces, realizando un procesamiento secuencial, se obtiene una nueva representación de vectores embebidos:

$$\mathbf{F}^{\mathbf{L}}_{\mathbf{I}} = \mathbf{\Phi}^{L}_{<1,2>}(\cdots \mathbf{\Phi}^{i}_{<1,2>}(\cdots \mathbf{\Phi}^{1}_{<1,2>}(\mathbf{I})))$$

En esta representación compuesta $\mathbf{F}^{\mathbf{L}}_{\mathbf{I}}$ es un conjunto de vectores embebidos que explotan la información visual, aprovechando la información contextual de múltiples módulos de *swin transformers* $\mathbf{\Phi}^i_{<1,2>}$. Como se detalla en la figura 7-f, el bloque BTS utiliza una representación de auto-atención, definida como:

$$MSA = 4hwC^2 + \left\{ \begin{array}{c} a = 2(hw)^2C \\ b = 2 \times 7^2 hwC \end{array} \right\}$$

En la cual se selecciona el componente a y b para el bloque $\mathbf{\Phi}^i_{<1>}$, $\mathbf{\Phi}^i_{<2>}$, respectivamente. Cabe destacar, además que en la aplicación secuencial de los módulos $\mathbf{\Phi}^i_{<1,2>}$, previamente se realiza un proceso de fusión parcial de los parches, aprovechando nuevas relaciones regionales (en la figura 7-FL, se ilustra este proceso), donde a cada vector embebido se le concatena información parcial del vector consecutivo.

En cuanto al texto, la sentencia *"A lung nodule in the imagen"* y la palabra*"nodule"* se codificaron como una representación constante F_T. Cabe aclarar que en este trabajo, se fijo esta representación unicamente a la sentencia mencionada, para poder explorar las relaciones

visuales extraídas desde el modelo fundacional. Básicamente, esta sentencia es proyectada a una representación $[\mathbf{Vt} \cup \mathbf{V_p} \cup \mathbf{V_s}]$, la cual brinda información codificada asociada con el token $\mathbf{t}$, la posición relativa de cada token $\mathbf{p}$ y una codificación relativa al segmento de la frase, en la cual está posicionada cada palabra $\mathbf{s}$, respectivamente. Esta representación embebida, con información contextual tanto del token como de la posición relativa de la frase y del segmento, se procesa a través de múltiples módulos de auto-atención (figura 7-f-BT), obteniendo una representación $\mathbf{F_T^L} = \mathbf{\Psi}^L(\cdots \mathbf{\Psi}^i(\cdots \mathbf{\Psi}^1(\mathbf{T})))$. Esta formulación sigue el modelamiento definido en la parte visual, como se puede observar en la figura 7-a.

Seguido, el conjunto de vectores embebidos, obtenidos de las representaciones visuales y de texto $(\mathbf{F_I}, \mathbf{F_T})$, son refinados a través de un bloque de procesamiento conjunto de atención, así obteniendo las características $(\mathbf{F_I''}, \mathbf{F_T''})$. Particularmente, este bloque de refinamiento de características hace uso de esquemas de auto-atención (SA), atención cruzada (CA) y atención deformable (DA) [58]. Primero, los vectores de texto $\mathbf{F_T^L}$ se mapea a un modulo de auto-atención $\mathrm{F}'_T = \mathrm{SA}(Q_{F_T}, K_{F_T}, V_{F_T}) = \mathrm{softmax}\left(\frac{Q_{F_T} K_{F_T}^{\dagger}}{\sqrt{d_k}}\right) V_{F_T}$, obteniendo vectores F'_T que aprovechan relaciones distantes en la misma representación. la arquitectura BERT implementada en esta rama de procesamiento de texto fue pre-entrenada con tareas no supervisada de modelado de lenguaje enmascarado (MLM) y la predicción de la siguiente oración (NSP). En MLM, se enmascaran aleatoriamente algunos tokens de la entrada y el modelo debe predecirlos basándose en el contexto circundante, fomentando una comprensión bidireccional del contexto. En NSP, el modelo recibe pares de oraciones y debe determinar si la segunda oración sigue a la primera en T, mejorando su capacidad para comprender relaciones entre oraciones.

Por otra parte, los vectores asociados a la representación visual $\mathbf{F_I^L}$ se mapean a un esquema de atención deformable (DA), que permite obtener relaciones visuales aprovechando esquemas de auto-atención, pero corregidos por campos vectoriales, denominados *offset*. Una ilustración

---

[58] Zhuofan XIA et al. "Vision transformer with deformable attention". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4794–4803.

de este módulo se presenta en el sub-cuadro c, de la figura 7. Particularmente, en este modelos de atención deformable se aprenden mapas de vectores de desplazamiento R'_P, a partir de una representación Υ, que corrigen la localización de las representaciones visuales, causada por el procesamiento convolucional. Entonces, las características de entrada se proyectan convolucionalmente a la arquitectura ofsset ($Q_F = F_I \cdot W_q$) y se mapean a la red $\Upsilon(Q_F)$, la cual actualiza las características $\hat{F}$, como: $\hat{F} \leftarrow R'_P(\Delta P)$, donde $R'_P = \Upsilon(Q_F) + R_P$. Desde esta representación corregida se realiza el proceso de multi-atención, realizando las proyecciones complementarias $K_{\hat{F}} = \hat{F} \cdot W_K$ y $V_{\hat{F}} = \hat{F} \cdot W_V$. Así cada modulo de atención es definido como $Z^m = \sigma\left(\frac{Q_F^m \cdot K_{\hat{F}}^m}{\sqrt{d}}\right) \cdot V_{\hat{F}}^m$ y la representación multicabeza conjunta es: $DA = \text{concat}(Z^1, Z^2, \ldots, Z^m)$.. Este modulo DA ha reportado mayor efectividad para representar objeto pequeños, capturando así dependencias contextuales internas.

Luego, en el mismo bloque de refinamiento de características se aplican bloques de atención cruzada utilizando las características intermedias, tanto de imagen como de texto ($\mathbf{F''_I}, \mathbf{F''_T}$). Estos bloques de atención cruzada se ejecutan de forma consecutiva, como:

$$\mathbf{F''_I} = \text{CA}(Q_{F'_I}, K_{F_T}, V_{F_T}) = \text{softmax}\left(\frac{Q_{F'_I} K_{F_T}^{\dagger}}{\sqrt{d_k}}\right) V_{F_T}$$

donde $\mathbf{F}\text{I}'$ es la representación visual refinada mediante la atención deformable, y $\mathbf{F''_I}$ es la representación visual actualizada con información textual a través de la atención cruzada. Análogamente, para la representación textual:

$$\mathbf{F''_T} = \text{CA}(Q_{F'_T}, K_{F''_I}, V_{F''_I}) = \text{softmax}\left(\frac{Q_{F'_T} K_{F''_T}^{\dagger}}{\sqrt{d_k}}\right) V_{F'_T}$$

donde $\mathbf{F'_T}$ es la representación textual refinada mediante auto-atención, y $\mathbf{F''_t}$ es la representación textual actualizada con información visual a través de la atención cruzada.

A continuación, estas características actualizadas pasan por el módulo de selección de *queries* guiados por lenguaje, cuyo objetivo es seleccionar las $\mathbf{F''_I}$ que mejor coincidan con las representaciones contextuales de $\mathbf{F''_T}$ mediante un proceso de comparación basado en máximos.

Este proceso permite la identificación de correspondencias de alta similitud entre las modalidades, utilizando un enfoque de selección de características con operaciones de máximos favoreciendo aquellas que estén alineadas en ambas modalidades. Formalmente de manera determinada se seleccionan 900 queries de la siguiente manera:

$$\mathbf{N_i} = TOP_{900}(Max(-1)(F_I'' F_T''^{\dagger}))$$

Entonces, en esta etapa el conjunto de tres vectores embebidos resultante ($\mathbf{F_I''}, \mathbf{F_T''}, \hat{\mathbf{F}}_\mathbf{I}{}''$) entran a una última etapa de decodificación multimodal para predecir las coordenadas de las cajas delimitadoras (*bounding boxes*) de los objetos detectados. Una ilustración de este módulo se encuentra en la figura 7-sección e. Este decodificador utiliza inicialmente un modelo de autoatención para los vectores que provienen de la selección de queries $\hat{\mathbf{F}}_\mathbf{I}'' = \mathrm{SA}(Q_{\hat{\mathbf{F}}_T''}, K_{\hat{\mathbf{F}}_T''}, V_{\hat{\mathbf{F}}_T''})$. Estos vectores son integrados con $\mathbf{F_I''}$, mediante un módulo de atención cruzada, aprovechando la redundancia de información, como: $\mathbf{N_I} = \mathrm{CA}(Q_{\hat{F}_I''}, K_{F_I''}, V_{F_I''})$ Estos vectores $\mathbf{N_i'}$ a su vez son integrados con la representación textual $\mathbf{F_T''}$, siguiendo también un modelo de atención cruzada $\hat{\mathbf{N}}_\mathbf{i}' = \mathrm{CA}(Q_{N_i}, K_{F_T''}, V_{F_T''})$. Los *queries*, ahora enriquecidas con información de ambas modalidades, se utilizan para predecir las coordenadas de las cajas delimitadoras de los objetos detectados. Una capa de proyección lineal (FFN) transforma estas representaciones en $\mathbf{N_i''}$ correspondientes a parámetros que definen las cajas, generando la respectiva correspondencia entre las características multimodales y las posiciones espaciales de cada candidato a nódulo $\mathbf{C_n} = \mathbf{I}(x_n, y_n, w_n, h_n)$.

4.3. Reducción de falsos positivos mediante un modelo de grafos con representaciones multi-escala

Los métodos computacionales para la detección de NP en imágenes de TC enfrenta múltiples desafíos, siendo uno de los más relevantes la abundancia de falsos positivos (FP) en sus salidas. Este problema surge debido a la similitud visual entre los nódulos y otras estructuras

anatómicas (como vasos sanguíneos y ramificaciones bronquiales) o artefactos (causados por movimientos del paciente o defectos en los detectores). Para abordar este reto, en este trabajo se implementó un enfoque basado en grafos con representaciones multi-escala que permite identificar las características locales y contextuales propias de los NP, descartando así, candidatos correspondientes a FPs.

En la figura 8 se muestra un esquema general de la arquitectura propuesta para la reducción de FPs. En primer lugar, se implementa un proceso de contextualización local sobre cada uno de los candidatos $\mathbf{C_n}$, obtenidos por el detector fundacional, expresado en la sección anterior. Para esto, las coordenadas de las cajas delimitadoras se dilatan mediante un factor $\rho \leq 1$, definido como un porcentaje relativo al tamaño de la caja original. Esto se representa como $\mathbf{C'_n} = \text{dilate}(\mathbf{C_n}, \rho) = \mathbf{I}(x_n, y_n, w_n(1+\rho), h_n(1+\rho))$, donde $\mathbf{C'_n}$ corresponde a las regiones dilatadas. De esta manera se garantiza que las regiones capturen suficiente contexto anatómico al rededor del NP, facilitando así su diferenciación con respecto a otras estructuras. Posteriormente, los parches extraídos son redimensionados uniformemente a un mismo tamaño $[H', W', C]$, donde H' y W' son las dimensiones espaciales, y C corresponde al número de canales. Este paso asegura una consistencia en la entrada para la siguiente etapa del procesamiento.

A continuación, se extraen características profundas de cada uno de los parches $\mathbf{C'_n}$, obtenidos en el paso anterior, con el objetivo de usarlas para la descripción local y contextual posterior. Para esto se implementó la red EfficientNet[59], la cual consiste en una arquitectura convolucional construida a partir de bloques MBConv (Mobile Inverted Bottleneck Convolution). Este tipo de convoluciones descomponen el proceso convencional en una convolución que actúa de manera independiente sobre cada canal de la entrada, seguida de una convolución que combina la información de todos los canales. Esto reduce la complejidad computacional

---

[59] Mingxing TAN and Quoc LE. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

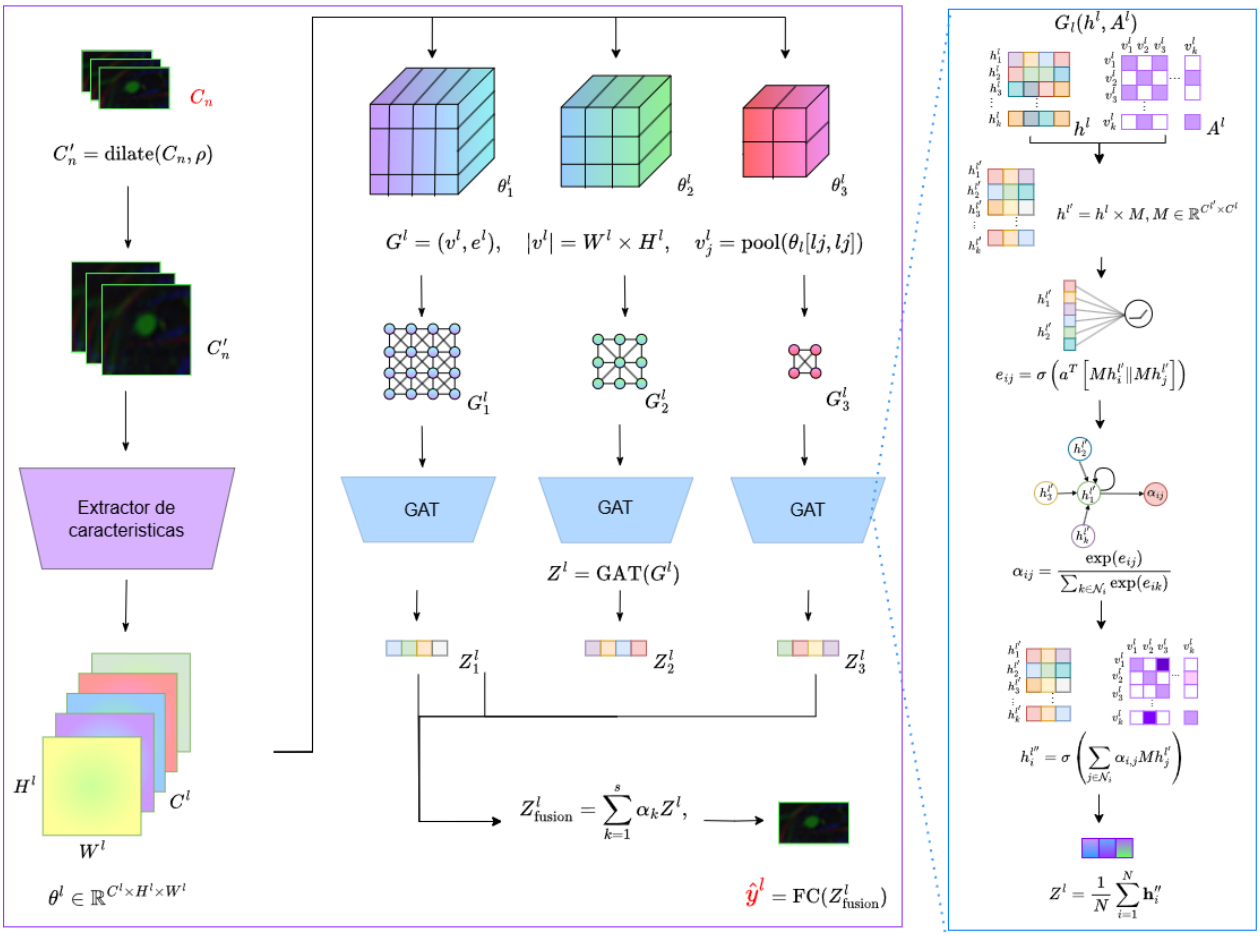

Figura 8. Esquema del RFP propuesto. El proceso comienza con la dilatación y reescalado de los nódulos candidatos (C_n) para incluir contexto adicional. A partir del parche dilatado, se extraen características (θ), las cuales se procesan a múltiples escalas (θ^l). Cada escala es transformada en un grafo (G^l) donde los nodos representan celdas de una cuadrícula sobre las características. Los grafos son procesados por GATs independientes, generando representaciones (Z^l) que se combinan en una fusión ponderada (Z^l_{fusion}) usando pesos entrenables. Finalmente, el vector fusionado pasa por una capa completamente conectada para obtener la predicción binaria ($\hat{y}^l$).

sin comprometer la capacidad del modelo para extraer características relevantes. Además, EfficientNet utiliza un enfoque de escalado uniforme que ajusta parámetros clave, como la profundidad de la red (número de capas) y el ancho de las capas (número de filtros o canales). Este enfoque facilita modificar el tamaño y la configuración del modelo, ofreciendo variantes como EfficientNet-B0, B1, B2, hasta B7, que se adaptan a diferentes necesidades en térmi-

nos de capacidad y recursos computacionales. Específicamente, en este trabajo se utilizó la versión de menor tamaño, EfficientNet-B0, debido a que su arquitectura ligera y eficiente es ideal para trabajar con conjuntos de datos limitados, como es el caso de las imágenes médicas de NP, reduciendo el riesgo de sobreajuste y optimizando el entrenamiento en condiciones de datos escasos. Desde esta arquitectura se extrajeron características de 3 niveles de profundidad diferentes de la red, con el objetivo de capturar información multiescala y de distinta complejidad semántica. Las características de las primeras capas de la red se utilizan para describir patrones de bajo nivel, como bordes y texturas de los candidatos a NP, mientras que las capas intermedias aportan representaciones de estructuras más complejas, tales como combinaciones de formas y patrones locales. Finalmente, las capas profundas proporcionan descripciones de alto nivel que capturan relaciones semánticas y contextuales globales dentro de cada parche $\mathbf{C'_n}$. Formalmente, cada capa l del modelo genera un conjunto de características $\theta^l \in \mathbb{R}^{C^l \times H^l \times W^l}$, donde C^l es el número de canales, y (H^l, W^l) son las dimensiones espaciales de las características en l. Esta estrategia multi-escala permite capturar información complementaria esencial para distinguir entre nódulos y estructuras no patológicas.

Posteriormente, las características extraídas en cada escala son utilizadas para representar grafos que modelen relaciones locales y contextuales de los parches. Con este fin, se aplana cada volumen de características $\theta^l \in \mathbb{R}^{C^l \times H^l \times W^l}$ obteniendo una matriz $\hat{\theta}^l \in \mathbb{R}^{C^l \times (H^l \cdot W^l)}$. Cada fila i de $\hat{\theta}^l$ representa un vector de características $h_i^l \in \mathbb{R}^{C^l}$ que describe al nodo v_i^l. En cuanto a las aristas e^l del grafo, estas se definen en función de la adyacencia espacial entre los pixeles, de tal forma que cada nodo, exceptuando a los de los bordes, tiene 8 nodos adyacentes. Teniendo en cuenta lo anterior, el grafo obtenido para la capa l se representa como $G^l = (v^l, e^l)$, donde $|v^l| = W^l \cdot H^l$.

Para el procesamiento de los grafos obtenidos se implementó una arquitectura profunda basada grafos, que incluye mecanismos de atención (conocida comunmente como *Graph*

Attention Network (GAT)), debido a su capacidad para asignar pesos adaptativos a las conexiones entre nodos, lo que permite capturar relaciones locales y contextuales de manera más efectiva. A partir de cada h_i^l, el modelo GAT aprende una nueva representación para cada nodo v_i^l, mediante la combinación de información proveniente de sus vecinos directos $\mathcal{N}(v_i^l)$. Este proceso puede describirse mediante la ecuación $h_i^{l'} = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{(i,j)} M h_j^l\right)$, donde $M \in \mathbb{R}^{d\times C^l}$ es la matriz de pesos que transforma el vector de características de tamaño C^l a uno de dimensión d, σ es una función de activación no lineal (como ReLU) y $\alpha_{(i,j)}$ representa el coeficiente de atención entre los nodos i y j, obtenido tras aplicar un mecanismo que evalúa la relevancia de cada vecino y denotado como $\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^\top [Mh_i^l \parallel Mh_j^l]))}{\sum_{k\in\mathcal{N}_i} \exp(\text{LeakyReLU}(\mathbf{a}^\top [Mh_i^l \parallel Mh_k^l]))}$, con $\parallel$ denotando la concatenación de los vectores y $\mathbf{a} \in \mathbb{R}^{2d}$ representando un vector entrenable que evalúa la interacción entre los nodos. De esta manera, la GAT refina iterativamente las representaciones de los nodos al resaltar las interacciones más significativas en el grafo. Extendiendo este proceso, la nueva representación de un nodo se obtiene al combinar la información de todos sus vecinos ponderada por los coeficientes de atención de la forma $h_i^{l''} = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{i,j} M h_j^{l'}\right)$. Al procesar todas las capas de la GAT, se requiere obtener una única representación que capture la información global del grafo, especialmente en tareas donde el objetivo es predecir atributos a nivel de grafo (como la clasificación). Para ello, se utiliza un mecanismo de *global pooling*, que consolida las representaciones finales de los nodos en un vector único. Este pooling se realiza mediante un promedio de las representaciones finales de todos los nodos en el grafo representado como $Z^l = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i''$ donde N es el número total de nodos en el grafo y $\mathbf{h}_i''$ corresponde a la última representación refinada del nodo i.

Para integrar la información proveniente de múltiples escalas, se realiza una fusión ponderada de las representaciones. Esto se logra asignando un peso entrenable α_k a cada escala y combinándolas mediante: $Z^l_{\text{fusion}} = \sum_{k=1}^{s} \alpha_k Z^l$, donde α_k está sujeto a la restricción: $\sum_{k=1}^{s} \alpha_k = 1, \quad \alpha_k \geq 0$. La representación fusionada Z^l_{fusion} encapsula tanto las relaciones

locales como globales, capturando patrones espaciales relevantes en distintas representaciones. Finalmente, la predicción para cada ROI se genera mediante una capa completamente conectada (*Fully Connected*, FC), que toma como entrada la representación fusionada: $\hat{y}^l = \mathrm{FC}(Z^l_{\mathrm{fusion}})$.

5. DISEÑO EXPERIMENTAL

5.1. Conjuntos de datos

La estrategia desarrollada fue ajustada y validada con el conjunto de datos públicos LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) de imágenes radiológicas TC del tórax. Este conjunto de imágenes consta de 1,018 estudios con un total de 2,742 nódulos. Además de las imágenes, el conjunto de datos incluye información y marcaciones relevantes relacionadas con la ubicación, características de malignidad y caracterización de los NP. Estas anotaciones fueron realizadas de forma independiente por hasta cuatro radiólogos expertos [60]. Entre las características definidas para los NP se encuentran el tamaño y la dificultad de detección de los NP según su densidad. La tabla 1 resume la distribución del conjunto de datos para entrenamiento, validación y prueba.

Tabla 1. Distribución del conjunto de datos

Conjunto	**Porcentaje**	**Scans**	**Imágenes**	**NP**
Entrenamiento	70%	582	1415	1534
Validación	10%	162	386	493
Prueba	20%	85	203	225

En cuanto al tratamiento de los datos para permitir que las imágenes radiológicas sean recibidas correctamente por el modelo, se extrajeron los cortes (slices) de cada volumen de estudio, las imágenes son convertidas a png y pre-procesadas con el enfoque de contextualización volumétrica para posteriormente redimensionar cada corte a 512 $\times$ 512 píxeles.

---

[60] 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". In: *Medical physics* 38.2 (2011), pp. 915–931.

5.2. Configuración de la arquitectura

El trabajo propuesto consiste en una red fundacional que procesa y extrae regiones que tienen una potencial asociación de NP y una red de refinamiento que define los NP. A continuación se detallan los parámetros y configuraciones necesarias:

- Para la red fundacional, se utilizó el modelo Grounding-DINO con parámetros ya predefinidos por la red, tales como la aplicación de escalas de aumento de datos que varían entre 480 y 800 píxeles, con un tamaño máximo de 1333 píxeles. Las escalas de redimensionamiento se establecieron en 400, 500 y 600 píxeles, mientras que las escalas de recorte fueron de 384 y 600 píxeles, asegurando que no hubiera solapamiento entre ellas. El tamaño del lote se fijó en 3, y se empleó el codificador de imágenes Swin-T 224 1k. Se utilizaron embebidos posicionales del tipo 'sine' con una temperatura de 20. La arquitectura del transformer incluyó 6 capas tanto en el codificador como en el decodificador, sin normalización previa. La dimensión de la capa 'feedforward' se estableció en 2048, y la dimensión de las capas ocultas en 256, sin aplicar dropout. Se configuraron 8 cabezas de atención y 900 'queries' entrenables con una dimensión de 4. Los puntos de atención en el codificador y decodificador se establecieron en 4. Se adoptó una estrategia de dos etapas estándar, sin compartir embebidos de las cajas delimitadoras de los candidatos a NP. La función de activación del transformer fue 'relu', compartiendo la predicción de embebidos de las cajas delimitadoras en el decodificador. Como metódo de regularización se introdujo ruido en las cajas delimitadoras con una escala de 1.0 y una proporción de ruido en las etiquetas de 0.5. La longitud máxima del texto se estableció en 256, utilizando el codificador de texto "bert-base-uncased" con las entradas "nodule" y "A lung nodule in the image". El optimizador seleccionado fue AdamW con una tasa de aprendizaje base de 0.0001 y un 'weight decay' de 0.0001, asignando una tasa de aprendizaje de 1e-05 para ambos codificadores. El entrenamiento se programó para 150 épocas, con una reducción de la tasa de aprendizaje a partir de la

cuarta época, guardando puntos de control en cada época.

- Para el reductor de falsos positivos, se utilizó EfficientNet-B0 como extractor de características. EfficientNet-B0 fue seleccionada de manera empírica tras realizar una evaluación comparativa frente a otros extractores de características del estado del arte (VGG19, ConvNext, DenseNet, ResNet50), donde esta red presentó un desempeño superior al integrarse con el modelo de grafos. Esta red se utilizó en tres niveles: nivel 3 ($[\text{batch}, 40, 28, 28]$), nivel 5 ($[\text{batch}, 112, 14, 14]$) y nivel 7 ($[\text{batch}, 320, 7, 7]$). Las características extraídas se transformaron en nodos con dimensiones $[\text{batch}, 784, 40]$, $[\text{batch}, 196, 112]$ y $[\text{batch}, 49, 320]$, respectivamente, conectados en grafos regulares mediante matrices de adyacencia. Se implementaron GATs para cada nivel, configuradas con dos capas: la primera con canales ocultos (32, 64 y 128) y cabezas de atención (1, 2 y 3), la segunda con canales de salida (16, 32 y 64). Las salidas de las GATs se proyectaron a 32 canales y se fusionaron con pesos aprendibles normalizados. Finalmente, se utilizó una capa lineal para la clasificación binaria. El entrenamiento se realizó durante 100 épocas con un tamaño de lote de 32, optimizador SGD (tasa de aprendizaje inicial 0.0001), y dropout de 0.3 en las capas GAT.

5.3. Validación

Para evaluar la efectividad del metodo propuesto, los conjuntos de datos fueron validados mediante tareas de detección y clasificación de candidatos potenciales a ser NP. Siendo la primera de ellas aplicada para el modelo fundacional antes y después de pasar por el RFP y la segunda exclusivamente dentro del RFP. Para la tarea de detección, la validación incluye métricas de amplio uso reportadas en el estado del arte, las cuales permiten cuantificar el numero de predicciones correctas comparado la cantidad de falsos positivos por imagen y por estudio completo TC (*scan*) y la precisión promedio de las detecciones. Las métricas usadas se detallan a continuación:

Precisión Media Promedio (mAP@IOU) La métrica mAP@IOU se utiliza para evaluar la precisión del modelo en la detección de NP con un umbral de Intersección sobre Unión (IoU). Esta métrica mide la capacidad del modelo para detectar NP con una superposición definida con las etiquetas en el conjunto de datos de prueba. La precisión media promedio se calcula de la siguiente manera:

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

donde k es el número de clases evaluados y AP_i es la precisión media para la i-ésima clase. Aquí $p(r)$ representa la precisión en un nivel de sensibilidad r. Este valor se obtiene calculando el área bajo la curva Precision-sensibilidad, que refleja la relación entre la proporción de verdaderos positivos sobre el total de predicciones positivas y la proporción de verdaderos positivos sobre el total de elementos reales positivos a diferentes umbrales de confianza. Una vez calculadas las AP para todas las clases evaluadas se promedian y de esta forma se obtiene el mAP donde un valor más alto indica un mejor rendimiento del modelo, reflejando una mayor precisión y sensibilidad en la detección de objetos.

Competition Performance Metric (CPM) y curva FROC CPM es una medida que evalúa la sensibilidad del modelo para diferentes cantidades de FP por imagen o por scan (conjunto de imagenes). Los diferentes valores de FP y de sensibilidad se obtienen al variar el umbral de confianza de las predicciones, lo que a su vez permite considerar más o menos predicciones como verdaderos positivos para el cálculos de las métricas. Luego, se promedian los valores de sensibilidad obtenidos para varias cantidades específicas de FP (normalmente siete) . Formalmente, esto se puede expresar como: $\text{CPM} = \sum_{c=1}^{7} \text{r}(FP_c)$, donde c representa cada uno de los niveles predefinidos de falsos positivos por imagen o scan (por ejemplo, 0.125, 0.25, 0.5, 1, 2, 4, 8), y $\text{r}(FP_c)$ es la sensibilidad del modelo al permitir FP_c falsos positivos. Este enfoque facilita la evaluación integral del rendimiento del modelo, mostrando qué tan bien mantiene una alta sensibilidad, mientras se controla la generación de FP. La

curva FROC (Free-Response Receiver Operating Characteristic) se utiliza para representar gráficamente esta relación entre sensibilidad y falsos positivos por imagen o scan. Esta curva ilustra el desempeño del modelo a lo largo de los distintos niveles de umbral, permitiendo una visualización clara de la compensación entre sensibilidad y generación de FP.

Esquema de validación Para evaluar el rendimiento del modelo de detección de NP, se llevaron a cabo análisis tanto a nivel de imagen como de *scan*, como se describe a continuación. Respecto al enfoque de imagen, se seleccionaron exclusivamente las imágenes del conjunto de prueba que contenían NP. Este enfoque permitió centrarse en la capacidad real del modelo para detectar y localizar nódulos, reduciendo el efecto de imágenes sin NP. De esta manera, se obtuvo una medición del rendimiento del modelo en escenarios donde está confirmada la presencia de NP.
Siguiendo esta validación a nivel de imagen, en primer lugar, se midió el aporte de cada una de las etapas del preprocesamiento, es decir, la segmentación del parénquima y la contextualización volumétrica, en términos de CPM, utilizando un umbral de IOU=0.25. Una vez validado el aporte del preprocesamiento propuesto, se seleccionó la mejor representación de datos y se validó nuevamente utilizando umbrales de IoU de 0.25, 0.5 y 0.75, para determinar la influencia del valor de umbral en el desempeño reportado. Un IoU más alto exige una mayor coincidencia entre la predicción del modelo y la etiqueta real, lo que representa un desafío mayor para el modelo, y por consiguiente, valores de CPM y de mAP menores.

Por otra parte, en la validación a nivel de scan, la mayoría de imágenes de prueba no contienen NP, representando un entorno clínico más realista, donde los nódulos son poco frecuentes. En este enfoque por scan, se midió la capacidad del modelo fundacional para aprender características nodulares con diferentes cantidades de datos de entrenamiento. Para esto se redujo progresivamente el porcentaje de datos de entrenamiento (usando el 100%, 80%, 60%, 40% y 20%) y el 100% del conjunto de prueba.
Posteriormente, se evaluó el aporte y desempeño del método RFP. Para esto, se calcularon las

métricas de exactitud (EXA), Precisión (PRE), Sensibilidad (SEN) y Puntaje F1 (F1). Estas métricas, utilizadas en la evaluación de modelos de clasificación binaria, permiten cuantificar la capacidad del modelo para distinguir entre nódulos y no nódulos. Un valor alto de EXA indica un alto porcentaje de clasificaciones correctas, mientras que PRE y SEN evalúan la capacidad del modelo para minimizar los falsos positivos y falsos negativos, respectivamente. El F1, siendo la media armónica de PRE y SEN, proporciona una medida balanceada del rendimiento general del modelo. Adicionalmente, se llevó a cabo una validación estratificada para medir la capacidad del modelo en la caracterización de los NP. Para esto se consideró diversos atributos de los NP, tales como su tamaño, textura, margenes y la dificultad que presentan para su detección por parte de expertos radiólogos. Estos atributos están directamente asociados a los problemas clínicos en la detección de NP y permiten comparar el rendimiento del modelo con la capacidad de detección humana demostrando así el potencial del modelo como herramiente para el soporte diagnóstico.

6. EVALUACIÓN Y RESULTADOS

El enfoque propuesto fue validado bajo dos contextos de detección (imagen - scan) con el fin de determinar el aporte del método a la hora de estimar la localización de NP en secuencias de TC. Así mismo también se validó un modulo desarrollado como reductor de falsos positivos y se realizó un estudio exhaustivo que comparó la capacidad de localización con características radiológicas de los nódulos. Las siguientes secciones presentan esta evaluación.

6.1. Caracterización por imagen

Tabla 2. Resultados de CPM para las diferentes etapas del procesamiento. Se muestra el valor de sensibilidad para diferentes cantidades de FP por imagen junto al valor promedio (CPM), considerando un umbral de IOU=0.25 entre la predicción y la etiqueta.

Tipo imagen	**1/8**	**1/4**	**1/2**	**1**	**2**	**4**	**8**	**CPM@0,25**
Imagen TC sin procesar	0.73	0.78	0.82	0.84	0.87	0.89	0.92	0.841 ± 0.061
Parénquima segmentado	0.74	0.82	0.84	0.86	0.87	0.91	0.92	0.857 ± 0.057
Contextualización volumétrica	**0.81**	**0.85**	**0.89**	**0.92**	**0.94**	**0.96**	**0.96**	**0.91 ± 0.052**

Para evaluar el impacto de las diferentes etapas del preprocesamiento propuesto, inicialmente se fijó el umbral de IOU=0.25 (establecido en la literatura como un umbral apropiado para objetos pequeños), y así se realizó el cálculo del CPM. Cabe destacar que en esta evaluación por componentes de la arquitectura se utilizaron imágenes crudas, imágenes con el parénquima segmentado y se analizó el aporte de la contextualización volumétrica. La Tabla 2 muestra los resultados obtenidos para este experimento. El uso de la imagen TC sin preprocesamiento muestra un rendimiento inferior, lo que sugiere una dificultad mayor para la arquitectura para explorar y determinar las regiones asociadas al nódulo, las cuales pueden ser etiquetadas erróneamente con objetos de fondo, ruido y otros segmentos en el pulmón (CPM=0.841). Sin embargo, al aplicar la segmentación del parénquima pulmonar se acota el área de interés para que el modelo solo procese la información espacial relevante, de esta forma

se evidencia una pequeña mejora (CPM=0.857). Por otra parte, se logra ver que cuantitativamente el enfoque de contextualización volumétrica propuesto supera considerablemente a los dos métodos anteriores, alcanzando un CPM de 0.91. Esto prueba la efectividad de la contextualización volumétrica para capturar diferencias entre nódulos y otras estructuras pulmonares.

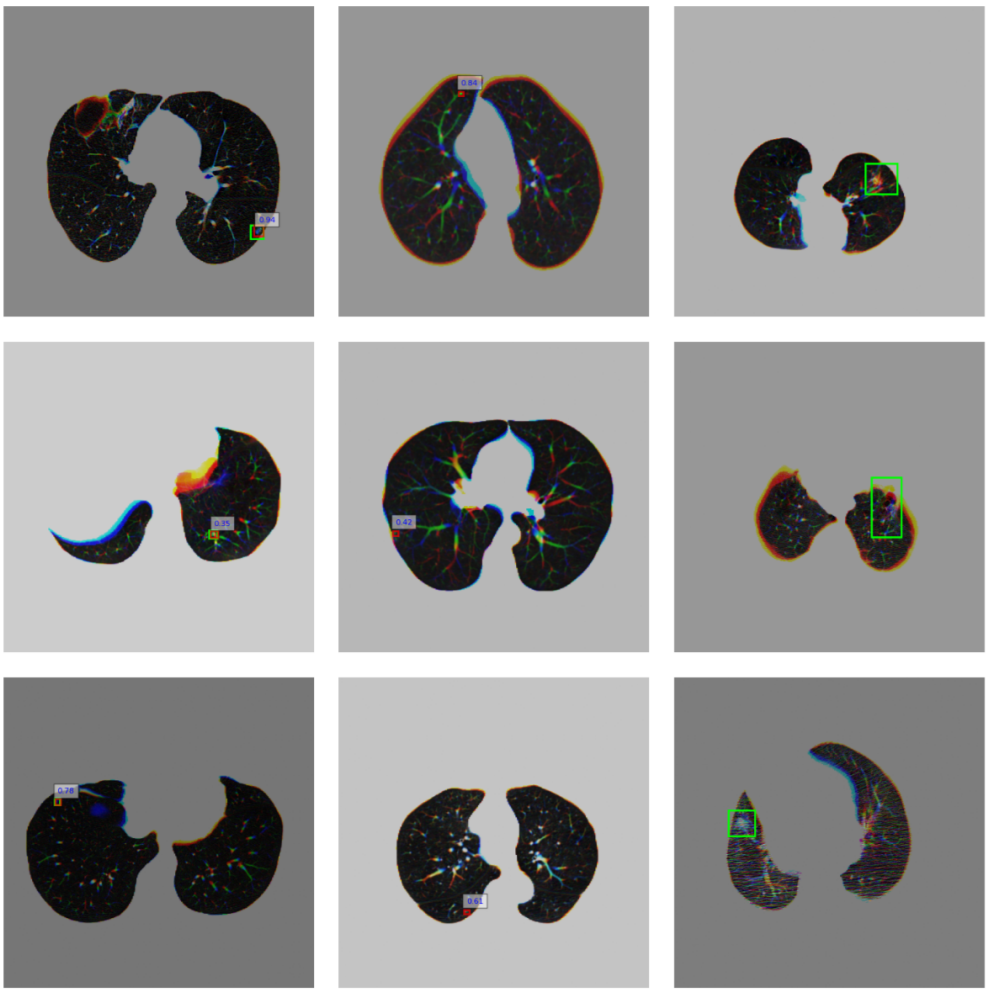

Figura 9. Resultados observacionales de la red de detección de candidatos a NP, donde a) representa predicciones exitosas sin ruido adicional, b) casos donde la red realiza predicciones sin acertar a ninguna etiqueta (FP) y c) casos donde el modelo no realizó predicciones y habían etiquetas (FN).

Observacionalmente, en la figura 9-a se pueden destacar ejemplos donde la red de candidatos

demuestra una capacidad notable para detectar NP, independientemente de su tamaño y ubicación. A pesar del buen rendimiento general en la detección de NP, se observa la aparición de una cantidad considerable de FP (ver Figura 9-b). Esta situación es especialmente notable en las imágenes TC donde no hay presencia de NP, lo que indica la necesidad de implementar métodos adicionales para reducir la tasa de FP y mejorar la precisión del modelo. Además, la Figura 9-c ilustra casos en los que la red de candidatos no detecta el NP presente, generando falsos negativos (FN). Esto subraya la importancia de equilibrar la sensibilidad y la precisión del modelo, optimizando la detección de NP y minimizando los FP.

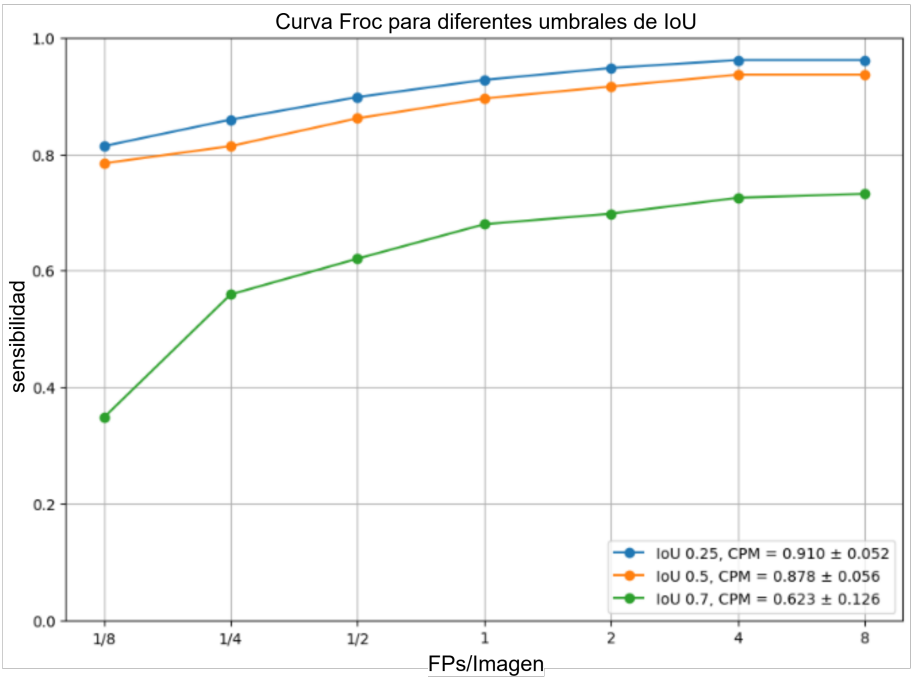

Figura 10. Curvas FROC con diferentes umbrales de IoU, donde la línea azul representa el rendimiento con una mayor permisividad al usar un IoU=0.25, la línea naranja muestra el rendimiento con un IoU=0.5, valor comúnmente utilizado en imágenes naturales y, la línea verde indica el rendimiento con un umbral más estricto de IoU=0.75.

Una vez analizado los tipos de entrada y mostrando la ventaja de una representación volumétrica, se procedió a evaluar el modelo con distintos umbrales IoU (0.25, 0.5 y 0.75). Los resultados mostrados en la figura 10 indican un rendimiento elevado en los tres umbrales, pasando de

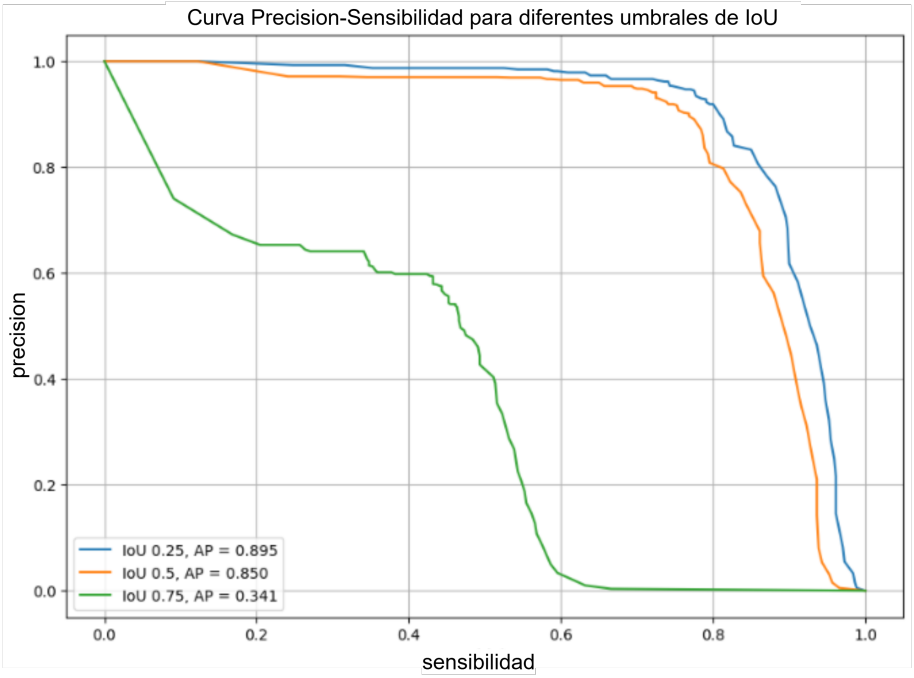

Figura 11. Curvas Precisión-Sensibilidad con diferentes umbrales de IoU: 0.25, 0.5 y 0.75, representadas por la linea azul, naranja y verde respectivamente.

0.91 a 0.87 y a 0.62, lo que representa un cambio de 0.04 y 0.29 respectivamente. Esta consistencia, particularmente entre los umbrales 0.25 (usado en objetos pequeños como NP) y 0.5 (usado en imágenes naturales), sugiere que el modelo mantiene una robustez significativa en la detección de NP, a pesar de tratarse de objetos considerablemente pequeños.

Adicionalmente, se calculó el mAP utilizando los mismos umbrales de IoU. Los resultados, mostrados en la figura 11, reflejan una alta precisión en la detección de NP en los umbrales 0.25 y 0.5, pasando de 0.895 a 0.85. Mientras que, para el umbral 0.75, se obtuvo un valor de mAP de 0.341, lo que indica una disminución significativa en el desempeño del modelo al aumentar drásticamente la restricción en la superposición requerida para considerar una detección como correcta. Esto sugiere que, aunque el modelo tiene un buen rendimiento general en detecciones menos restrictivas, es necesario mejorar su capacidad para localizar los NP con mayor precisión en escenarios donde se exige una mayor exactitud espacial.

6.2. Caracterización por volumen

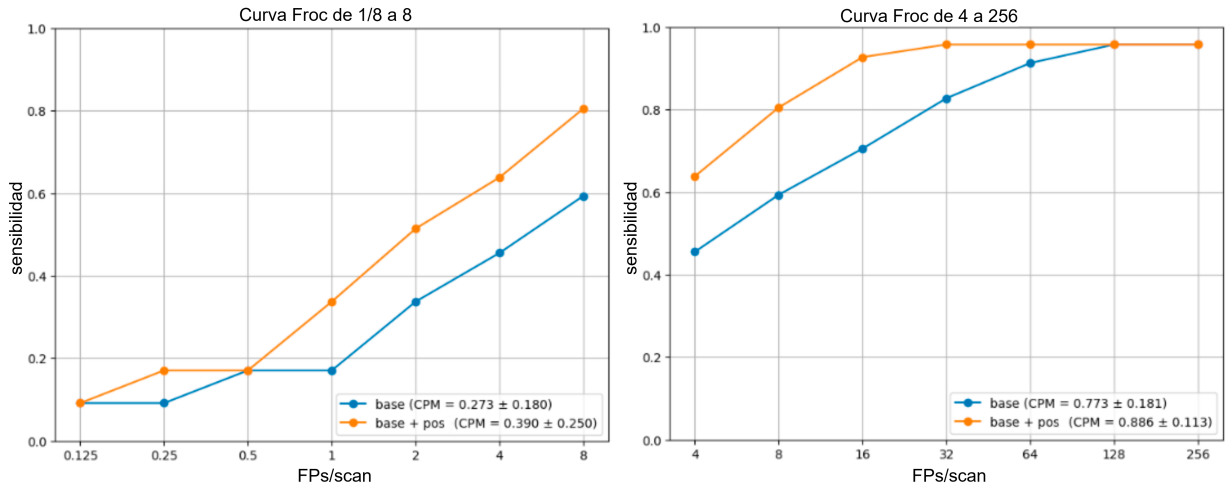

Figura 12. Curvas FROC comparativas entre las configuraciones del modelo de detección base y con el posprocesamiento de caracterización por volumen. La gráfica de la izquierda muestra los resultados en el rango de 1/8 a 8 FP/scan, mientras que la gráfica de la derecha presenta los resultados en el rango de 4 a 256 FP/scan.

Para emular la evaluación del modelo en contextos clínicos, en este trabajo se validó el enfoque propuesto incluyendo estudios completos de tomografía. En este contexto, la herramienta debe sugerirle al experto posibles candidatos a nódulos a lo largo de todo el scan. Para realizar esta evaluación, se decidió entonces calcular el CPM por volumen (Ver figura 12, linea azul). Considerando que las predicciones obtenidas corresponden a cajas delimitadoras 2D y que pueden obtenerse más de una predicción por nódulo, en diferentes slices del scan, se implementó un proceso de agrupamiento basado en superposición espacial 3D para asociar las predicciones de diferentes slices. Para hacer esta evaluación, se agruparon las predicciones por objeto, es decir si en un slice subsiguiente se encuentra otra predicción con un IOU superior a 0.5 se determina que es una predicción del mismo objeto por lo que se agrupan y se eligen las coordenadas de la predicción con mayor confianza (Ver figura 12, linea naranja). Tras este proceso, se observa que el rendimiento máximo alcanzado en el esquema por imagen, con 128 FP por scan, mejora al agruparse las predicciones, logrando reducirse a 64 FP por scan.

Este rendimiento corresponde al nivel que anteriormente se alcanzaba con 4 FP por imagen.

Tabla 3. Resultados de CPM en el conjunto de datos de prueba para diferentes porcentajes de entrenamiento. Se muestra el valor de sensibilidad para diferentes cantidades de FP por scan junto al valor promedio (CPM), considerando un umbral de IOU=0.25 entre la predicción y la etiqueta.

Porcentaje de entrenamiento	**CPM@0.25 [1/8 - 8]**	**CPM@0.25 [4 - 256]**
20	0.44	0.83
40	**0.46**	0.86
60	0.44	0.85
80	0.42	0.85
100	0.42	**0.87**

Seguido, en una evaluación experimental posterior, se evaluó la capacidad de la aproximación fundacional en escenarios con datos limitados. Este escenario también se tuvo en cuenta, considerando los contextos reales, donde la adquisición y etiquetado de imágenes puede tener limitaciones marcadas, lo cual conduce a que las aproximaciones puedan fallar. Para esta evaluación, se dividió el conjunto de entrenamiento original en fracciones del 100%, 80%, 60%, 40% y 20%, correspondientes a 1393, 1115, 835, 556, y 277 imágenes, respectivamente. En la Tabla 3 se resumen los resultados obtenidos en este experimento. Tras entrenar y evaluar los modelos con cada subconjunto, se observó que el rendimiento se mantenía prácticamente constante obteniendo resultados de $\text{CPM}_{1/8:8}$ respectivamente de [0.42, 0.42, 0.44, 0.46 ,0.44] y [0.87, 0.85, 0.85, 0.86, 0.83] para $\text{CPM}_{4:256}$, siendo la mayor diferencia de unicamente 0.025, incluso con las reducciones significativas en la cantidad de datos de entrenamiento.

Este comportamiento sugiere que la aproximación fundacional implementada posee una notable capacidad para ajustarse eficazmente, incluso cuando se dispone de cantidades limitadas de datos. Tal característica es especialmente valiosa en este contexto, donde la recopilación de datos etiquetados es costosa y laboriosa. Estos hallazgos están en línea con estudios previos que indican que los modelos de gran escala pueden comportarse bien, incluso en regímenes de datos pequeños, sin embargo, el problema de los FP es persistente. Lo anterior puede estar asociado a el entrenamiento realizado y aprovechado de grandes volúmenes de datos públicos, en escenarios naturales y, sin limitaciones de privacidad. Además de los componentes basados

en módulos de atención, los cuales están diseñados para extraer y retener patrones complejos de los datos anteriormente mencionados.

6.3. Reductor de falsos positivos - RFP

A pesar de los resultados sobresalientes del modelo fundacional y de su destacada detección de NP, estos resultados siguen siendo dramáticamente permisivos con la generación de falsos positivos. Es por ello, que en la metodología propuesta se incluyó un esquema de procesamiento posterior para hacer esta reducción, aprovechando la contextualización brindada por una red basada en grafos. Para ello, en este trabajo se tomaron regiones predichas por el modelo fundacional, las cuales correspondían a falsos y verdaderos positivos. En particular, se utilizaron 16,888 parches como conjunto de entrenamiento, distribuidos igualitariamente entre las clases positiva y negativa (50% cada una). De manera similar, el conjunto de validación estuvo compuesto por 2,356 parches, manteniendo la misma proporción de clases. Por último, el conjunto de prueba incluyó 2,421 parches de la clase positiva y 19,472 de la clase negativa.

Para la selección de regiones que servirían como entrada al reductor en entrenamiento y validación, se seleccionaron las predicciones del modelo fundacional con un valor de confianza superior a 0.3, priorizando las de la clase positiva para garantizar una distribución igualitaria entre clases. Mientras que el conjunto de prueba utilizó la totalidad de las predicciones realizadas por Grounding-DINO en su propio conjunto de prueba. Todos los parches obtenidos fueron procesados y redimensionados a un tamaño uniforme de 224×224 píxeles para garantizar consistencia en el análisis posterior.

Para evaluar este RFP propuesto y medir la contribución de una representación con grafos, inicialmente se evaluaron modelos convolucionales clásicos en esta tarea. Seguido, estos modelos fueron comparados con representaciones de grafos que utilizaban como extractor de características diferentes esquemas convolucionales. La Tabla 4 presenta una comparación entre modelos como VGG19, ConvNext, DenseNet, ResNet50 y EfficientNet, tanto en su

forma base como al integrar grafos. Como se puede observar la arquitectura construida con grafos y características extraídas desde la EfficientNet superó a todas las combinaciones, alcanzando una EXA de 0.85, una PRE de 0.42, una SEN de 0.87, y un F1 de 0.57. Estos resultados destacan la eficacia del modelamiento con grafos para representar los NP y mejorar la discriminación frente a los FP.

Tabla 4. Resultados de rendimiento comparativo entre diferentes modelos con y sin el uso de grafos. Se presentan las métricas EXA, PRE, SEN y F1 para cada modelo, tanto en su forma base como al integrar grafos.

Método	**ACC**	**PREC**	**REC**	**F1**
Vgg19	0.79	0.33	0.83	0.47
ConvNext	0.82	0.37	0.82	0.51
DenseNet	0.82	0.37	**0.90**	0.53
RestNet50	0.83	0.39	0.85	0.53
EfficientNet	0.79	0.32	0.78	0.45
Vgg19 + Grafos	0.84	0.39	0.81	0.53
ConvNext + Grafos	0.82	0.36	0.84	0.51
DenseNet + Grafos	0.85	0.42	0.84	0.56
RestNet50 + Grafos	0.83	0.39	0.87	0.54
EfficientNet + Grafos	**0.85**	**0.42**	0.87	**0.57**

Entonces, considerando las características profundas obtenidas desde la EfficientNet se prosiguió a una evaluación de diferentes configuraciones, basadas en grafos, para maximizar la sensibilidad mientras se limitaba el número de falsos positivos por scan (FP/scan). Como linea base se consideró el modelo fundacional sin RPF y también una versión de este modelo pero con posprocesamiento (Pos). La Tabla 5 resume los resultados obtenidos con las distintas configuraciones, evaluadas en términos del coeficiente CPM para intervalos de FP de 1/8 a 8 y de 4 a 256, así como la cantidad de FP/scan en la que se alcanza la sensibilidad máxima. El diseño del método propuesto incluyó varias configuraciones exploratorias. Una de ellas fue el enfoque por *consenso*, donde las tres salidas binarias de cada modelo GAT se unificaron mediante votación por clase, es decir, la clase final seleccionada corresponde a aquella predicha por al menos dos de los tres grafos. Este método logró un valor de $\text{CPM}_{1/8:8} = 0.436$, un $\text{CPM}_{4:256} = 0.85$ y alcanzó su valor máximo de sensibilidad con una baja cantidad de FP (8).

Tabla 5. Resultados comparativos de las configuraciones del RFP y MF evaluadas en términos de CPM y FP/scan. Se muestran los valores de CPM para intervalos de FP de 1/8 a 8 y de 4 a 256, así como la cantidad de FP/scan en los que se alcanza la sensibilidad máxima.

Método	**CPM$_{1/8:8}$**	**CPM$_{4:256}$**	**MAX FP/SCAN**
Consenso	0.436	0.85	8
Fusión ponderada	0.429	0.876	8
Fusión ponderada + Pos	**0.509**	0.73	**4**
—-	-	-	-
Fundacional sin RPF	0.273	0.773	128
Fundacional sin RPF + Pos	0.39	**0.886**	32

Otra configuración evaluada fue la de *"fusión ponderada"*, donde se asignaron pesos entrenables a las salidas de los modelos, combinándolas a través de una capa totalmente conectada para obtener una única predicción. Esta configuración presentó resultados similares a la estrategia anterior, alcanzando un valor de $\text{CPM}_{1/8:8} = 0.429$ y un $\text{CPM}_{4:256} = 0.876$.

De las configuraciones evaluadas, el método propuesto de "Fusión ponderada + Pos", en el cual se realizo un posprocesamiento a las predicciones de entrada mediante una agrupación volumétrica, demostró ser el más efectivo, logrando un $\text{CPM}_{1/8:8} = 0.509$ con solo 4 FP/scan. Este resultado es significativo en comparación con el modelo de detección "Grounding-DINO + Pos", que en el mismo intervalo (1/8 - 8) obtuvo únicamente 0.39, lo que subraya el impacto positivo del RFP. Sin embargo, en intervalos más permisivos con los FP (4-256), el método sin reductor alcanza un valor de CPM mayor de 0.886 en 32 FP/scan, mientras que el reductor llega en el mismo intervalo únicamente a 0.773. Esto indica que el reductor logra su cometido al equilibrar la cantidad de detecciones exitosas y los FP, aunque aún elimina una cantidad considerable de NP verdaderos.

6.4. Análisis de características radiológicas de NP

Una vez realizada la evaluación por componentes y la inclusión del reductor de falsos positivos (RFP), en esta sección, se reporta un análisis extra que relaciona la capacidad de detección de NP, según variables radiológicas como el tamaño, la textura, los márgenes y la sutileza (ver

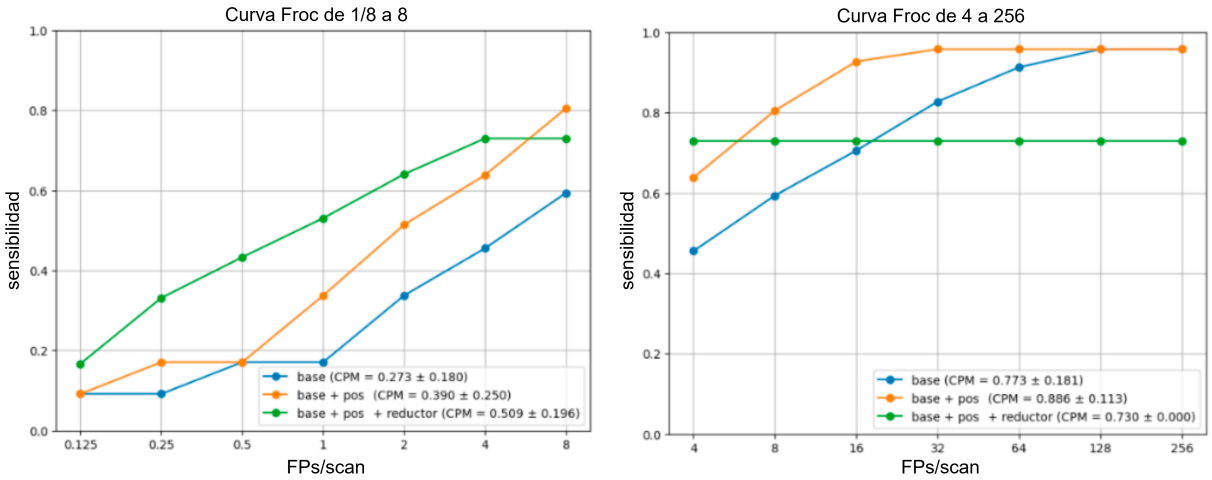

Figura 13. Curvas FROC comparativas entre las configuraciones del modelo de detección y el RFP. La gráfica de la izquierda muestra los resultados en el rango de 1/8 a 8 FP/scan, mientras que la gráfica de la derecha presenta los resultados en el rango de 4 a 256 FP/scan. Se observa que el reductor de falsos positivos propuesto logra un desempeño superior en el rango de bajo FP/scan, destacando su efectividad en escenarios de mayor precisión.

figura 14). Este análisis permite identificar fortalezas y limitaciones específicas del modelo en la detección de diferentes tipos de NP de manera estratificada, facilitando la interpretación de los resultados desde una perspectiva clínica.

Como se reporta en la figura 14-a, en un análisis estratificado por tamaño, el modelo alcanzó una tasa de detección del 100% para NP con tamaños superiores a 30 mm, 92% para aquellos entre 6 y 30 mm, y 74% para NP menores de 6 mm. Como era de esperarse, estos resultados demuestran que el rendimiento del modelo disminuye a medida que el tamaño del NP se reduce, evidenciando una limitación en la identificación de nódulos pequeños. Sin embargo, a pesar de esta tendencia, los resultados obtenidos para nódulos pequeños son prometedores, considerando este rango de tamaño. Este resultado es consistente con la literatura, donde se reporta que los nódulos menores de 6 mm presentan desafíos notables debido a su bajo contraste, su menor prominencia en el parénquima pulmonar y su propensión a confundirse con estructuras anatómicas normales. Aunque el rendimiento global del modelo es notable, esta resultado en particular señala una oportunidad de mejora, considerando la relevancia

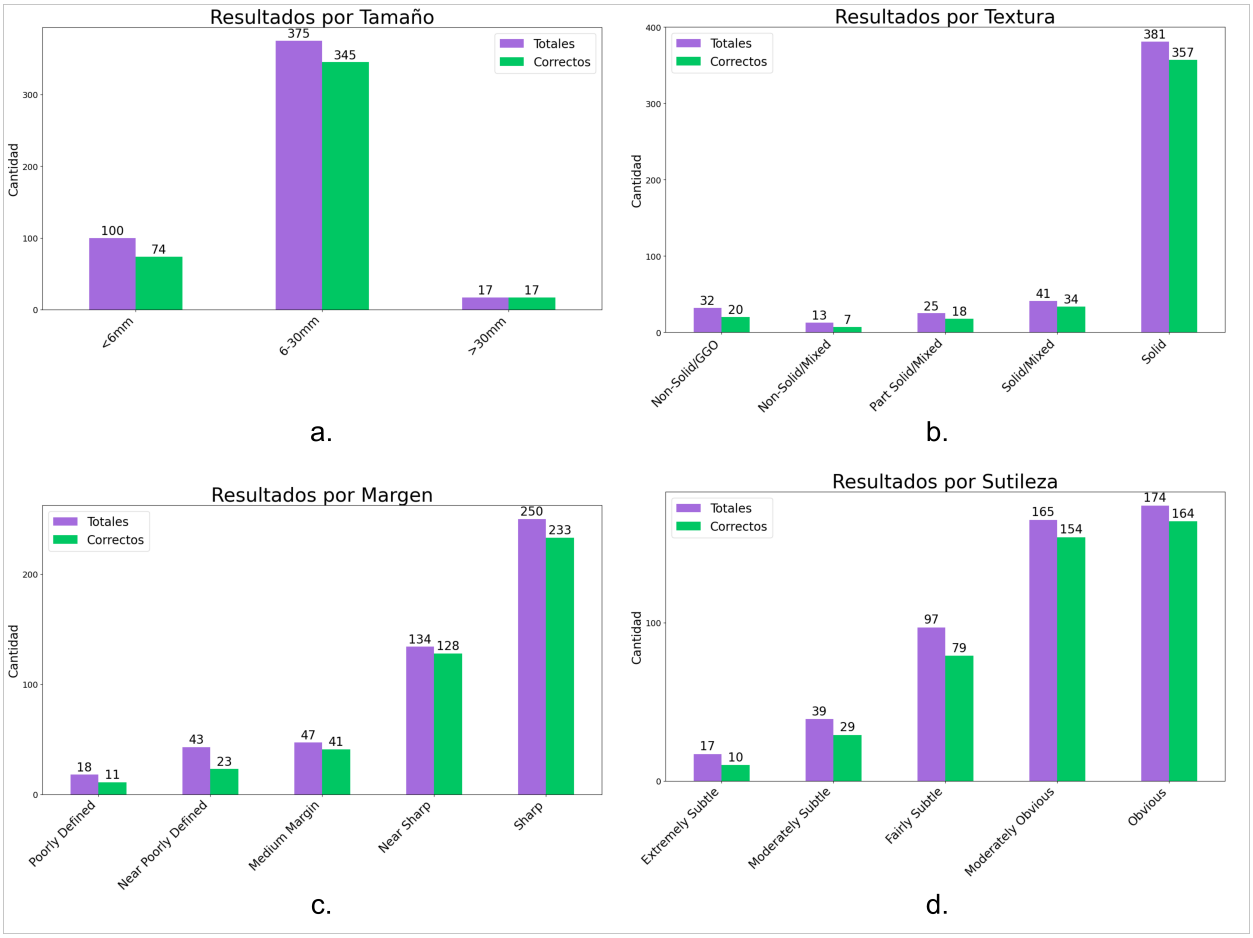

Figura 14. Resultados de caracterización de los NP, donde en la primera columna y primera fila se encuentra el diagrama de barras correspondiente al desempeño por tamaño, la segunda columna primera fila por por sutiliza o dificultad de detección según los radiologías, la primera columna segunda fila por margenes y la segunda columna en segunda fila al desempeño por textura

clínica de los nódulos pequeños en el diagnóstico temprano del cáncer de pulmón.

En la evaluación por textura, el modelo mostró un mejor desempeño en nódulos con una constitución sólida (ver figura 14-b). La precisión alcanzó un 93.7% para nódulos sólidos, disminuyendo a 82.9% para mixtos-sólidos, 72% para mixtos con parte sólida, 53.8% para mixtos sin parte sólida y 62.5% para nódulos de vidrio esmerilado. Los resultados para esta característica sugieren que la densidad del nódulo es un factor que influye significativamente en el rendimiento del modelo. Esto se explica por la relación entre la densidad del nódulo

y la intensidad en la imagen de TC, ya que los nódulos sólidos suelen presentar una mayor diferencia de intensidad con respecto al tejido pulmonar circundante, facilitando su detección. Por el contrario, los nódulos con baja densidad, como los de vidrio esmerilado, generan una menor diferencia de contraste, lo que dificulta su identificación precisa. Este comportamiento refleja la necesidad de enfoques más especializados para abordar este tipo de texturas con características menos densas.

En el análisis por márgenes, el modelo presentó un rendimiento notablemente alto para NP con márgenes bien definidos (93.2%) y casi bien definidos (95.5%), como se observa en la figura 14-c. No obstante, la precisión disminuyó significativamente en NP con márgenes poco definidos (61.1%) y mal definidos (53.5%). Estos resultados muestran que la claridad en la definición de los márgenes es un factor determinante para el desempeño del modelo. Los márgenes bien definidos permiten al modelo identificar contornos nítidos que facilitan la detección de los NP. En contraste, los márgenes poco o mal definidos suelen estar asociados a nódulos más complejos o a características morfológicas que dificultan la diferenciación entre el nódulo y el tejido circundante, generando una mayor cantidad de FN.

Por último, se realizó un análisis considerando la dificultad de detección percibida por los 4 radiólogos, también llamada "sutileza" (ver figura 14-d). El desempeño del modelo en función de la sutileza mostró tasas de detección del 94.2% para nódulos obvios, 93.3% para moderadamente obvios, 81.4% para sutiles, 74.3% para moderadamente sutiles y 58.8% para extremadamente sutiles. Este resultado subraya la capacidad del modelo para identificar incluso los NP más difíciles, con una tasa de detección superior al 50% en los casos extremadamente sutiles. Este hallazgo es especialmente significativo en el contexto clínico, ya que los nódulos más difíciles de detectar suelen ser los que pasan desapercibidos en la práctica diaria y tienen implicaciones negativas directas en el diagnóstico temprano.

7. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo presentó una arquitectura fundacional que captura el contexto de los NP en imágenes de TC, considerando relaciones anatómicas y acoplando una representación geométrica para apoyar la reducción de FP durante la predicción. En particular, se implementó el modelo fundacional Grounding-DINO como detector de nódulos candidatos, y se diseñó una red de grafos multiescala para reducir los FP, esto con el objetivo de apoyar el diagnostico de CP. Este enfoque fue entrenado y validado utilizando el conjunto de datos LIDC-IDRI.

El modelo propuesto alcanzó un CPM por imagen de 0.91 superando trabajos del estado del arte que han reportado puntajes de 0.79 [61] y 0.88 [62], respectivamente. El trabajo propuesto reportó estos resultados en el LIDC, mientras que los otros trabajos utilizaron Luna16, un subconjunto de datos del LIDC que excluye nódulos pequeños y observaciones con sutileza baja [63]. El modelo propuesto, basado en un esquema fundacional, evidencia beneficios de generalización y aprovechamiento de aprendizaje en otros contextos, lo que facilita la personalización de herramientas en problemas específicos, como la detección de NP.

En cuanto a la evaluación de componentes de la metodología propuesta, cabe destacar la contextualización volumétrica que evidenció una ganancia de mas de 6% en el CPM, logrando brindar contextos volumétricos, en las observaciones. Así mismo, el modelo fue competitivo frente al uso de diferentes umbrales de IoU ([0.25, 0.5, 0.75]), mostrando robustez destacable

---

[61] Hongtao XIE et al. "Automated pulmonary nodule detection in CT images using deep convolutional neural networks". In: *Pattern recognition* 85 (2019), pp. 109–119.

[62] Xia HUANG et al. "Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks". In: *Computerized Medical Imaging and Graphics* 74 (2019), pp. 25–36.

[63] 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". In: *Medical image analysis* 42 (2017), pp. 1–13.

en IoU=0.5, pero también evidenciando ciertas limitaciones en el IoU=0.75, lo que puede prevenir su uso en ambientes críticos, como el apoyo a procedimientos quirúrgicos. El método fundacional implementado también fue adaptado para responder a predicciones volumétricas, logrando desde el agrupamiento tridimensional reducir predicciones redundantes sin afectar la sensibilidad máxima, pasando de 128 FP/scan a 32 FP/scan.

Unos de los principales intereses en la comunidad de imágenes diagnósticas es el uso de estas herramientas computacionales en ambientes con pocos datos, o contextos ampliamente desbalanceados. En este sentido el método propuesto mantuvo un rendimiento consistente incluso cuando se redujo significativamente el conjunto de entrenamiento (hasta el 20%), con una variación máxima de apenas 0.025 en su desempeño. Esto evidencia la capacidad de adaptabilidad y la robustez del método propuesto frente a escenarios con restricciones en la cantidad de datos disponibles. Además en la metodología propuesta se incluyó un modulo de reducción de falsos positivos, aprovechando la contextualización espacial y relaciones anatómicas, a través de esquemas basados en grafos. Este modulo RFP aumentó significativamente el desempeño del modelo, permitiendo una considerable reducción del FP/scan, alcanzando una sensibilidad máxima de 0.509 con tan solo 4 FP/scan. Esto contrasta notablemente con el modelo sin RFP, el cual obtuvo una sensibilidad máxima desde 32 FP/scan.

En este trabajo también se realizó una evaluación exhaustiva del esquema propuesto, incorporando un análisis de los resultados con respecto a las características radiológicas de los NP. Mostrando una mejor adaptación para detección de nódulos grandes y sólidos, mientras se mostró desventajas en los nódulos acotados como vidrio esmerilado. También se mostró una correlación de la capacidad de detección con la definición del margen y el grado de sutileza, observado en las TC. De hecho, a pesar de tener un desempeño menor en nódulos con baja sutileza, el modelo puede potencialmente beneficiar el soporte diagnóstico, debido a su localización en estos escenarios.

Como trabajo futuro se considera relevante evaluar el modelo con otros conjuntos de datos para determinar su capacidad de generalización, así como también explorar mecanismos al-

ternativos para mejorar la eficacia en la reducción de falsos positivos. En este sentido, la implementación de enfoques multimodales en los datos, como la integración de imágenes PET, podría enriquecer las representaciones contextuales y mejorar la detección de NP. También con los nuevos avances metodológicos se espera explorar codificación de instrucciones textuales que permitan guiar y ampliar la variabilidad de las representaciones visuales, logrando, como en los escenarios expertos, tener un soporte para una búsqueda y localización eficaz de los NP.

BIBLIOGRAFÍA

AHMADYAR, Yashar et al. "Hierarchical approach for pulmonary-nodule identification from CT images using YOLO model and a 3D neural network classifier". In: *Radiological physics and technology* 17.1 (2024), pp. 124–134 (cit. on p. 32).

ALBERT JEROME, S et al. "Watershed segmentation with CAFIS and RCNN classification for pulmonary nodule detection". In: *IETE Journal of Research* 69.8 (2023), pp. 5052–5063 (cit. on p. 31).

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". In: *Medical physics* 38.2 (2011), pp. 915–931 (cit. on p. 54).

BLANDIN KNIGHT, Sean et al. "Progress and prospects of early detection in lung cancer". In: *Open biology* 7.9 (2017), p. 170070 (cit. on p. 19).

CHENG, Tianheng et al. *YOLO-World: Real-Time Open-Vocabulary Object Detection*. 2024. arXiv: 2401.17270 [cs.CV] (cit. on p. 25).

DEL CIELLO, Annemilia et al. "Missed lung cancer: when, where, and why?" In: *Diagnostic and interventional radiology* 23.2 (2017), p. 118 (cit. on pp. 14, 16, 17, 36).

DOSOVITSKIY, Alexey et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV] (cit. on p. 18).

DUMA, Narjust; SANTANA-DAVILA, Rafael, and MOLINA, Julian R. "Non–small cell lung cancer: epidemiology, screening, diagnosis, and treatment". In: *Mayo Clinic Proceedings*. Vol. 94. 8. Elsevier. 2019, pp. 1623–1640 (cit. on pp. 14, 16, 17, 36).

FERLAY, Jacques et al. “Cancer statistics for the year 2020: An overview”. In: *International journal of cancer* 149.4 (2021), pp. 778–789 (cit. on pp. 14, 16, 17, 19, 36).

GEORGE, Jose; SKARIA, Shibon; VARUN, VV, et al. “Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans”. In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. SPIE. 2018, pp. 347–355 (cit. on pp. 15, 16).

GIRSHICK, Ross. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448 (cit. on p. 22).

GIRSHICK, Ross et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587 (cit. on p. 22).

GONG, Meiling et al. “A review of non-maximum suppression algorithms for deep learning target detection”. In: *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*. Vol. 11763. SPIE. 2021, pp. 821–828 (cit. on p. 22).

GU, Yu et al. “A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning”. In: *Computers in biology and medicine* 137 (2021), p. 104806 (cit. on p. 17).

GU, Zhongxuan et al. “Cross attention guided multi-scale feature fusion for false-positive reduction in pulmonary nodule detection”. In: *Computers in Biology and Medicine* 151 (2022), p. 106302 (cit. on p. 34).

HAMILTON, Will; YING, Zhitao, and LESKOVEC, Jure. “Inductive Representation Learning on Large Graphs”. In: *Advances in Neural Information Processing Systems*. Ed. by I. GUYON et al. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 29).

HAO, Rui; QIANG, Yan; YAN, Xiaofei, et al. “Juxta-vascular pulmonary nodule segmentation in PET-CT imaging based on an LBF active contour model with information entropy and joint vector”. In: *Computational and mathematical methods in medicine* 2018 (2018) (cit. on p. 20).

HOFMANNINGER, Johannes et al. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem”. In: *European Radiology Experimental* 4 (2020), pp. 1–13 (cit. on pp. 40, 41).

HU, Edward J et al. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021) (cit. on p. 26).

HUANG, Xia et al. “Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks”. In: *Computerized Medical Imaging and Graphics* 74 (2019), pp. 25–36 (cit. on p. 72).

IWANO, Shingo et al. “Computer-aided diagnosis: A shape classification of pulmonary nodules imaged by high-resolution CT”. In: *Computerized Medical Imaging and Graphics* 29.7 (2005), pp. 565–570. DOI: https://doi.org/10.1016/j.compmedimag.2005.04.009 (cit. on p. 19).

JAIN, Sweta; CHOUDHARI, Pruthviraj, and GOUR, Mahesh. “Pulmonary lung nodule detection from computed tomography images using two-stage convolutional neural network”. In: *The Computer Journal* 66.4 (2023), pp. 785–795 (cit. on p. 30).

KHEMANI, Bharti et al. “A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions”. In: *Journal of Big Data* 11.1 (2024), p. 18 (cit. on p. 27).

KIPF, Thomas N. and WELLING, Max. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG] (cit. on p. 28).

KIRILLOV, Alexander et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV] (cit. on p. 25).

LI, Bin et al. "Detection of pulmonary nodules in CT images based on fuzzy integrated active contour model and hybrid parametric mixture model". In: *Computational and mathematical methods in medicine* 2013 (2013) (cit. on p. 20).

LIANG, Jinglun et al. "Reducing False-Positives in Lung Nodules Detection Using Balanced Datasets". In: *Frontiers in Public Health* 9 (2021). DOI: 10.3389/fpubh.2021.671070 (cit. on p. 33).

LIN, T. "Focal Loss for Dense Object Detection". In: *arXiv preprint arXiv:1708.02002* (2017) (cit. on p. 24).

LIU, Kehong. "Stbi-yolo: A real-time object detection method for lung nodule recognition". In: *IEEE Access* 10 (2022), pp. 75385–75394 (cit. on p. 31).

LIU, Shilong et al. *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. 2024. arXiv: 2303.05499 [cs.CV] (cit. on pp. 18, 25, 42).

LIU, Wei et al. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37 (cit. on p. 23).

MA, He et al. "Automatic pulmonary ground-glass opacity nodules detection and classification based on 3D neural network". In: *Medical Physics* 49.4 (2022), pp. 2555–2569 (cit. on p. 41).

MAI, Juanyun et al. “Mhsnet: Multi-head and spatial attention network with false-positive reduction for lung nodule detection”. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2022, pp. 1108–1114 (cit. on p. 33).

MAMMERI, Selma et al. “Early detection and diagnosis of lung cancer using YOLO v7, and transfer learning”. In: *Multimedia Tools and Applications* 83.10 (2024), pp. 30965–30980 (cit. on p. 31).

MARINAKIS, Ioannis; KARAMPIDIS, Konstantinos, and PAPADOURAKIS, Giorgos. “Pulmonary Nodule Detection, Segmentation and Classification Using Deep Learning: A Comprehensive Literature Review”. In: *BioMedInformatics* 4.3 (2024), pp. 2043–2106 (cit. on p. 35).

MARTIN, Maria D. et al. “Lung-RADS: Pushing the Limits”. In: *RadioGraphics* 37.7 (2017). PMID: 29053407, pp. 1975–1993. DOI: 10.1148/rg.2017170051. eprint: https://doi.org/10.1148/rg.2017170051 (cit. on p. 19).

MAYNORD, Michael et al. “Semi-supervised training using cooperative labeling of weakly annotated data for nodule detection in chest CT”. In: *Medical Physics* 50.7 (2023), pp. 4255–4268 (cit. on p. 41).

MEI, Jie et al. “SANet: A slice-aware network for pulmonary nodule detection”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.8 (2021), pp. 4374–4387 (cit. on pp. 15, 16, 18).

METS, Onno M et al. “Incidental perifissural nodules on routine chest computed tomography: lung cancer or not?” In: *European radiology* 28 (2018), pp. 1095–1101 (cit. on p. 20).

MKINDU, Hassan; WU, Longwen, and ZHAO, Yaqin. “Lung nodule detection in chest CT images based on vision transformer network with Bayesian optimization”. In: *Biomedical Signal Processing and Control* 85 (2023), p. 104866 (cit. on pp. 18, 32).

MONKAM, Patrice et al. “Detection and classification of pulmonary nodules using convolutional neural networks: a survey”. In: *Ieee Access* 7 (2019), pp. 78075–78091 (cit. on pp. 15, 16, 20, 35).

ORGANIZATION, World Health. *Global cancer burden growing, amidst mounting need for services*. News release. World Health Organization. Feb. 2024. URL: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services (cit. on p. 36).

RADFORD, Alec et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV] (cit. on p. 25).

RAMESH, Aditya et al. “Hierarchical text-conditional image generation with clip latents”. In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), p. 3 (cit. on p. 25).

RAMEZANI, Hooman; ALEMAN, Dionne, and LÉTOURNEAU, Daniel. “Lung-DETR: Deformable Detection Transformer for Sparse Lung Nodule Anomaly Detection”. In: *arXiv preprint arXiv:2409.05200* (2024) (cit. on p. 32).

REDMON, Joseph et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788 (cit. on p. 23).

EL-REGAILY, Salsabil Amin et al. “Multi-view Convolutional Neural Network for lung nodule false positive reduction”. In: *Expert systems with applications* 162 (2020), p. 113017 (cit. on p. 33).

REN, Shaoqing et al. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149 (cit. on p. 22).

RUBIN, Geoffrey D. “Lung nodule and cancer detection in computed tomography screening”. In: *Journal of thoracic imaging* 30.2 (2015), pp. 130–138 (cit. on pp. 14, 16, 17, 19, 36).

SANCHEZ-LENGELING, Benjamin et al. “A Gentle Introduction to Graph Neural Networks”. In: *Distill* (2021). https://distill.pub/2021/gnn-intro. DOI: 10.23915/distill.00033 (cit. on p. 28).

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”. In: *Medical image analysis* 42 (2017), pp. 1–13 (cit. on p. 72).

SHEN, Zhiqiang et al. “WS-LungNet: A two-stage weakly-supervised lung cancer detection and diagnosis network”. In: *Computers in Biology and Medicine* 154 (2023), p. 106587. DOI: https://doi.org/10.1016/j.compbiomed.2023.106587 (cit. on p. 30).

SIM, Yongsik et al. “Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs”. In: *Radiology* 294.1 (2020), pp. 199–209 (cit. on p. 17).

SU, Ying; LI, Dan, and CHEN, Xiaodong. “Lung nodule detection based on faster R-CNN framework”. In: *Computer Methods and Programs in Biomedicine* 200 (2021), p. 105866 (cit. on p. 30).

SUN, Lingma et al. “Attention-embedded complementary-stream CNN for false positive reduction in pulmonary nodule detection”. In: *Computers in Biology and Medicine* 133 (2021), p. 104357 (cit. on p. 34).

TAN, Mingxing and LE, Quoc. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114 (cit. on p. 49).

TAN, Mingxing; PANG, Ruoming, and LE, Quoc V. "Efficientdet: Scalable and efficient object detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10781–10790 (cit. on p. 24).

THANOON, Mohammad A. et al. "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images". In: *Diagnostics* 13.16 (2023). DOI: 10.3390/diagnostics13162617 (cit. on p. 35).

VAROQUAUX, Gaël and CHEPLYGINA, Veronika. "Machine learning for medical imaging: methodological failures and recommendations for the future". In: *NPJ digital medicine* 5.1 (2022), p. 48 (cit. on pp. 15, 16).

VELIČKOVIĆ, Petar et al. *Graph Attention Networks*. 2018. arXiv: 1710.10903 [stat.ML] (cit. on p. 29).

WU, Xiaosheng et al. "YOLO-MSRF for lung nodule detection". In: *Biomedical Signal Processing and Control* 94 (2024), p. 106318 (cit. on p. 31).

XIA, Zhuofan et al. "Vision transformer with deformable attention". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4794–4803 (cit. on p. 46).

XIE, Hongtao et al. "Automated pulmonary nodule detection in CT images using deep convolutional neural networks". In: *Pattern recognition* 85 (2019), pp. 109–119 (cit. on p. 72).

XU, Jing et al. “An improved faster R-CNN algorithm for assisted detection of lung nodules”. In: *Computers In Biology And Medicine* 153 (2023), p. 106470 (cit. on p. 30).

XU, Rui et al. “Sgda: towards 3d universal pulmonary nodule detection via slice grouped domain attention”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2023) (cit. on p. 41).

ZAIDI, Syed Sahil Abbas et al. “A survey of modern deep learning based object detection models”. In: *Digital Signal Processing* 126 (2022), p. 103514 (cit. on p. 21).

ZHANG, Lin et al. “Graph neural networks for image-guided disease diagnosis: A review”. In: *iRADIOLOGY* 1.2 (2023), pp. 151–166 (cit. on p. 35).

ZHANG, Shengyu et al. “Instruction tuning for large language models: A survey”. In: *arXiv preprint arXiv:2308.10792* (2023) (cit. on p. 26).

ZHENG, Sunyi et al. “Automatic Pulmonary Nodule Detection in CT Scans Using Convolutional Neural Networks Based on Maximum Intensity Projection”. In: *IEEE Transactions on Medical Imaging* 39.3 (2020), pp. 797–805. DOI: 10.1109/TMI.2019.2935553 (cit. on p. 34).

ZHOU, Jie et al. “Graph neural networks: A review of methods and applications”. In: *AI Open* 1 (2020), pp. 57–81. DOI: https://doi.org/10.1016/j.aiopen.2021.01.001 (cit. on p. 27).

ZHU, Hongbo et al. “MR-forest: a deep decision framework for false positive reduction in pulmonary nodule detection”. In: *IEEE Journal of Biomedical and Health Informatics* 24.6 (2019), pp. 1652–1663 (cit. on p. 33).