Universidad Internacional de La Rioja
Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Inteligencia artificial

Diagnóstico multiclase de enfermedades pulmonares en radiografías con CNN destacando zonas relevantes con Grad-CAM

Trabajo fin de estudio presentado por:	Sebastián Cadena Molina Nelson Chicaiza Barahona
Tipo de trabajo:	Comparativa de Soluciones
Director/a:	Alejandro Cervantes Rovira
Fecha:	23/07/2025

Índice de contenidos

Índice de Figuras

Índice de tablas

Índice de Fórmulas

Resumen

Este trabajo propone un sistema de diagnóstico automatizado para enfermedades pulmonares mediante el uso de redes neuronales convolucionales (CNN) aplicadas a radiografías de tórax. El estudio se centra en cuatro patologías de alta incidencia clínica: cardiomegalia, neumotórax, engrosamiento pleural y nódulo pulmonar. Ante la carencia de radiólogos en entornos clínicos vulnerables, se busca ofrecer una herramienta que brinde soporte diagnóstico confiable y explicable.

Se utilizó el conjunto de datos público ChestX-ray14, que contiene más de 100,000 imágenes con etiquetas clínicas. Se aplicaron técnicas de preprocesamiento y etiquetado multiclase, seguidas de la implementación de tres arquitecturas CNN de referencia: DenseNet121, MobileNetV2 y ResNet50. Estas se entrenaron y evaluaron utilizando métricas como accuracy, recall y F1-score.

El estudio integra la técnica Grad-CAM para la generación de mapas de activación que permiten interpretar visualmente qué zonas de la imagen influyeron en la predicción del modelo. Esta funcionalidad es clave para incrementar la confianza clínica y facilitar la validación de los resultados.

Se llevó a cabo una comparativa exhaustiva entre las tres arquitecturas, evaluando su rendimiento cuantitativo y su capacidad de explicabilidad visual. DenseNet121 demostró un enfoque más preciso sobre regiones afectadas, mientras que ResNet50 ofreció un equilibrio entre precisión y activación visual coherente. MobileNetV2, aunque eficiente computacionalmente, mostró limitaciones en la focalización de algunas patologías.

Finalmente, se discutieron las limitaciones del sistema y se propusieron líneas futuras, como la integración de segmentación, el uso de etiquetas clínicas expertas y la validación en escenarios hospitalarios reales.

Palabras Clave: Diagnóstico médico asistido por IA, Radiografía de tórax, Redes neuronales convolucionales (CNN), Grad-CAM, Clasificación multiclase.

Abstract

This work presents the development of an automated diagnostic system for pulmonary diseases using convolutional neural networks (CNN) applied to chest X-ray images. The research focuses on four clinically significant pathologies: cardiomegaly, pneumothorax, pleural thickening, and pulmonary nodules. Motivated by the shortage of radiologists in vulnerable clinical settings, the goal is to provide a reliable and interpretable tool to support medical decision-making.

The public ChestX-ray14 dataset was used, comprising over 100,000 labeled radiographic images. The data underwent preprocessing, including resizing, normalization, and multilabel annotation. Three CNN architectures were implemented and compared: DenseNet121, MobileNetV2, and ResNet50. These models were trained and evaluated using performance metrics such as accuracy, recall, and F1-score.

To enhance interpretability, the Grad-CAM (Gradient-weighted Class Activation Mapping) technique was integrated. This method produces heatmaps overlaid on the original image, highlighting the regions that most influenced the model's prediction. This visualization enables healthcare professionals to understand and verify the model's diagnostic behavior.

A comparative analysis was conducted to evaluate the quantitative and qualitative performance of each model. DenseNet121 demonstrated superior precision and attention to localized features, while ResNet50 balanced predictive accuracy with coherent visual activation. MobileNetV2, although computationally efficient, exhibited less consistent focus in certain cases, particularly with subtle findings like nodules.

The study concludes by acknowledging limitations such as class imbalance and potential label noise. Future work includes incorporating segmentation techniques, expert-curated labels, multi-label classification, and clinical validation in real-world hospital environments.

This project contributes to the advancement of explainable AI in medical imaging, offering an accessible and effective tool for supporting diagnosis in low-resource healthcare systems.

Distribución y estructura de la memoria

Tabla 1. *Organización del trabajo en grupo.*

Organización del trabajo en grupo - Desarrollo de la memoria	
Apartado de la memoria	**Responsables**
Introducción	Javier Chicaiza, Sebastián Cadena
Contexto y estado del arte	Javier Chicaiza, Sebastián Cadena
Objetivos concretos y metodología de trabajo	Javier Chicaiza, Sebastián Cadena
Planteamiento de la comparativa	Javier Chicaiza, Sebastián Cadena
Desarrollo de la comparativa	Javier Chicaiza, Sebastián Cadena
Discusión y análisis de resultados	Javier Chicaiza, Sebastián Cadena

1. Introducción

1.1. Motivación

Las enfermedades pulmonares representan un desafío significativo continuo para la salud mundial, siendo una de las principales causas de mortalidad en todo el mundo. Cada año, patologías como la cardiomegalia, engrosamiento pleural, neumotórax y nódulo pulmonar, afectan a millones de personas, con una importancia clínica considerable, por lo que su diagnóstico oportuno es clave para evitar complicaciones severas y mejorar el pronóstico de los pacientes. (Candel et al., 2023).

A pesar de los avances en la medicina en el año 2021 murieron cerca de 2.1 millones de personas a causa de la neumonía, en los cuales más de 500.000 muertes fueron de niños menores a cinco años y más de un millón fueron adultos mayores de 70 (World Pneumonia Day, 2024).

En la actualidad, el diagnóstico de estas condiciones suele apoyarse en radiografías de tórax, sin embargo, el análisis e interpretación de las mismas están sujeto a la experiencia y respuesta de médico tratante, lo que condiciona a errores, demoras o interpretaciones inconsistentes (De la Cruz et al., 2021), además de factores como recursos o limitaciones a especialistas.

La Inteligencia Artificial y en particular las redes neuronales convolucionales (CNN) han destacado por su potencial en la clasificación de imágenes médicas, permitiendo reconocer patrones complejos con un alto nivel de precisión en comparación con médicos expertos, logrando analizar de forma eficiente una mayor cantidad de radiografías en un menor tiempo (Malek & Soufi, 2025).

El presente trabajo busca aportar con una solución accesible, replicable y viable para mejorar la detección automática de enfermedades pulmonares comunes, mediante el uso de inteligencia artificial, con un enfoque orientado a problemas actuales del ámbito médico y computacional.

1.2. Planteamiento del trabajo

El presente Trabajo de Fin de Máster (TFM) es el desarrollo de un modelo de clasificación multiclase de enfermedades pulmonares en radiografías de tórax, utilizando redes neuronales convolucionales aplicados a radiografías de tórax. El trabajo abordará en la clasificación multiclase de cinco condiciones importantes por su alta prevalencia y su visibilidad en estudios por imagen: cardiomegalia, engrosamiento pleural, neumotórax, nódulo pulmonar y normalidad.

Se integrará la técnica GRAD-CAM (Gradient-weighted Class Activation Mapping) con el objetivo de conseguir visualizar las áreas afectadas para el entrenamiento del modelo. El resultado del modelo no remplaza el diagnóstico de un médico a cargo, el modelo aporta un alto valor para la toma de decisiones y transparencia de resultados requerido para la aceptación del trabajo en entornos médicos (Georgakopoulou et al., 2024).

El modelo será entrenado y validado con el dataset público ChestX-ray14 con la finalidad de entregar un trabajo oportuno en el tiempo requerido sin incurrir en permisos sobre el uso de datos clínicos reales, asegurando la replicabilidad del estudio.

El trabajo busca comparar distintas arquitecturas de redes neuronales convolucionales (ResNet50, DenseNet121 y MobileNetV2) y evaluar su desempeño en una tarea de clasificación multiclase de enfermedades pulmonares, utilizando métricas como precisión, recall, F1-score y matriz de confusión. Además, se incorpora el uso de Grad-CAM como herramienta de explicabilidad para visualizar las regiones de las radiografías que influyen en las decisiones del modelo, permitiendo identificar la arquitectura más eficaz y eficiente para su implementación en entornos clínicos reales y de recursos limitados.

1.3. Estructura del trabajo

Este Trabajo Fin de Máster (TFM) está organizado de manera lógica y progresiva, permitiendo un análisis integral del problema abordado, la metodología utilizada y los hallazgos alcanzados. A continuación, se describe cada capítulo con mayor detalle:

Capítulo 1 – Introducción: En este capítulo se establece el contexto general del problema de investigación, subrayando la importancia del diagnóstico oportuno de enfermedades pulmonares mediante el uso de técnicas de inteligencia artificial. Se presentan los objetivos generales y específicos del trabajo, así como la motivación personal y científica que impulsó el estudio. Además, se justifica la elección de modelos de aprendizaje profundo aplicados a imágenes médicas y se anticipan los desafíos técnicos y éticos que conlleva su implementación. Finalmente, se incluye una vista previa de la estructura del documento.

Capítulo 2 – Estado del Arte: Aquí se realiza un análisis profundo de la literatura científica reciente relacionada con el uso de redes neuronales convolucionales (CNN) y mecanismos de atención en la clasificación de imágenes médicas, especialmente radiografías de tórax. Se examinan los fundamentos teóricos de los modelos DenseNet121, MobileNetV2 y ResNet50, así como la utilidad de herramientas de explicabilidad como Grad-CAM. El capítulo también contrasta diversos enfoques utilizados en investigaciones similares, estableciendo el marco teórico y técnico que sustenta el presente estudio.

Capítulo 3 – Metodología: Este capítulo describe detalladamente la estrategia experimental adoptada. Se presentan las características del dataset NIH ChestX-ray14, incluyendo su tamaño, tipo de patologías y limitaciones del etiquetado. Se explican los procesos de preprocesamiento aplicados a las imágenes, como normalización, cambio de tamaño y aumentos de datos. A continuación, se detalla la configuración técnica de los tres modelos analizados, los parámetros de entrenamiento, el uso de validación cruzada y las funciones de pérdida. También se explican las métricas de evaluación utilizadas (precisión, recall, F1-score, matriz de confusión) y la aplicación de Grad-CAM para análisis visual.

Capítulo 4 – Resultados: En este apartado se exponen los resultados obtenidos por cada modelo (DenseNet121, MobileNetV2 y ResNet50) aplicados a las cinco clases de estudio: cardiomegalia, engrosamiento pleural, nódulo, neumotórax y clase normal. Se presentan

comparativas cuantitativas a través de tablas de métricas y visualizaciones de las matrices de confusión, así como ejemplos concretos de interpretaciones Grad-CAM que revelan las zonas pulmonares donde cada modelo enfocó su atención. La sección también incluye el análisis de los errores más frecuentes, diferenciando entre verdaderos y falsos positivos y negativos para cada patología.

Capítulo 5 – Discusión: Este capítulo se dedica a interpretar los resultados expuestos anteriormente. Se analizan en profundidad las diferencias de rendimiento entre los modelos y se reflexiona sobre las posibles causas de los errores de clasificación. Se evalúa la capacidad de cada arquitectura para identificar las distintas enfermedades, destacando los casos en los que Grad-CAM contribuyó a una mejor comprensión del comportamiento del modelo. También se discute la aplicabilidad clínica de los hallazgos y las implicaciones de usar explicabilidad visual para respaldar decisiones médicas.

Capítulo 6 – Conclusiones y Trabajo Futuro: Finalmente, se resumen los principales hallazgos del estudio, destacando los aportes metodológicos, las fortalezas de los modelos empleados y las limitaciones encontradas. Se identifican las barreras actuales en cuanto al etiquetado, validación clínica, clasificación multiclase y balance de datos. A partir de estas limitaciones, se plantean líneas de trabajo futuro que incluyen mejoras en la calidad del dataset, incorporación de segmentación anatómica, implementación de clasificadores multietiqueta, validación con radiólogos expertos, integración en sistemas hospitalarios y consideraciones éticas para su adopción en entornos reales.

2. Contexto y estado del arte

2.1. Contexto del problema

La identificación de enfermedades pulmonares como la cardiomegalia, engrosamiento pleural, neumotórax y nódulo pulmonar, mediante imágenes de radiografía de tórax representa un desafío relevante para la salud global. A pesar de ser una técnica accesible y ampliamente utilizada, su interpretación está influenciada por diferencias en el juicio clínico, la disponibilidad de especialistas y la calidad de la imagen, lo cual puede derivar en errores en el diagnóstico, demoras en el tratamiento o decisiones clínicas inadecuadas.

En zonas con pocos recursos, donde hay escasez de radiólogos especializados o el trabajo es excesivo, estas dificultades se hacen aún más evidentes. En estos casos, la inteligencia artificial y en particular el aprendizaje profundo con redes neuronales llamadas CNN ha demostrado ser de gran ayuda para realizar diagnósticos con rapidez y precisión.

Sin embargo, la mayoría de los modelos desarrollados hasta la fecha se enfocan en tareas binarias (presencia/ausencia de una enfermedad específica), dejando de lado la complejidad clínica de escenarios multiclase en los que pueden coexistir distintas patologías. Además, se requieren métodos explicables que ayuden a los profesionales de la salud a confiar en la decisión del modelo, haciendo indispensable el uso de herramientas como Grad-CAM para la visualización de las regiones relevantes en la imagen.

2.2. Enfermedades Pulmonares

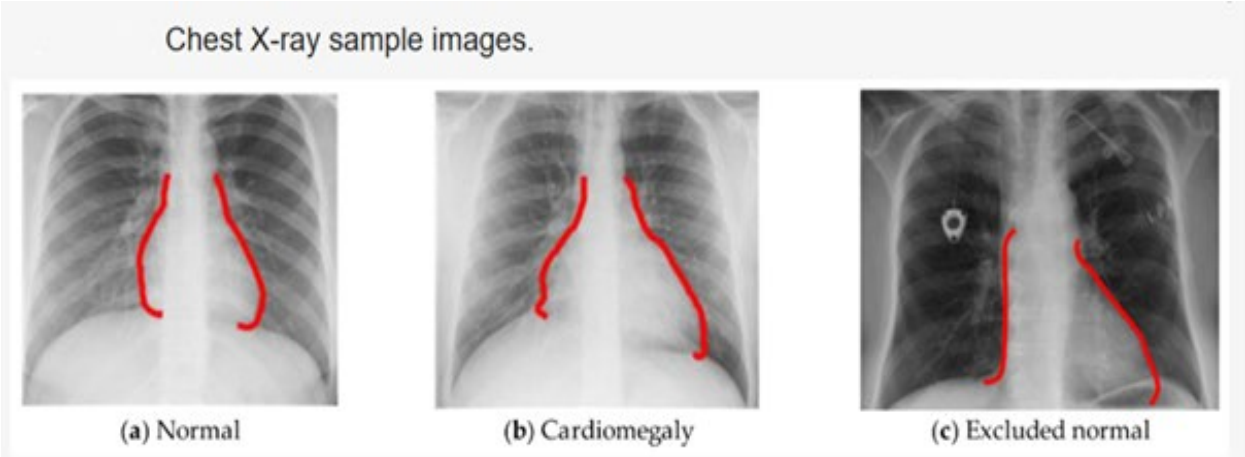
2.2.1. Cardiomegalia

Aproximadamente la mitad de las personas diagnosticadas con insuficiencia cardíaca fallecen en los 5 años siguientes al diagnóstico. La cardiomegalia se refiere al agrandamiento del corazón, se determina cardiomegalia cuando el diámetro transversal de la silueta cardíaca en una radiografía de tórax posterior o anterior es mayor o al 50% del diámetro transversal del tórax, también conocido como aumento del índice cardiotorácico. Se puede provocar esta enfermedad por hipertensión, cardiopatía isquémica, cardiopatía valvular y varios tipos de miocardiopatía (Amin & Siddiqui, 2025).

Las principales características de la cardiomegalia se muestran en la **Figura 1** mediante radiografías de tórax donde se puede evidenciar esta condición, esta condición se caracteriza por:

- Reduce el espacio torácico para la expansión pulmonar
- Puede provocar edema pulmonar por congestión venosa
- Se asocia con síntomas como ortopnea, disnea y paroxismos nocturnos.

Figura 1. *Imágenes de muestra de radiografía de tórax.*



Fuente: A Development and Validation of an AI Model for Cardiomegaly Detection in Chest X-rays

Entre los estudios que se han realizado con la cardiomegalia podemos mencionar los siguientes:

1. **Detección de cardiomegalia con U-Net personalizado y ChestX-ray8:** Este estudio se centró en la detección temprana de cardiomegalia usando un modelo híbrido de U-Net personalizado con transferencia de aprendizaje. Usaron el dataset público ChestX-ray8, del cual extrajeron imágenes etiquetadas de cardiomegalia y controles normales. Se realizó preprocesamiento: normalización, realce y reducción de resolución para ajustar a la arquitectura. El entrenamiento incluyó etapas de aumento (rotaciones, zoom, traslación) y se aplicó compresión de imágenes para optimizar tiempos sin perder información clínica. La evaluación se realizó con validación cruzada, obteniendo precisión del 94 %, sensibilidad 96.2 % y especificidad 92.5 %.

Se hizo comparación con otros modelos preentrenados, mostrando mejora medible. Además, integraron U-Net para segmentación del corazón antes de clasificación, lo que reduce ruido y mejora foco. El estudio discute la limitación del dataset (desequilibrio, etiquetas automáticas extraídas por NLP) como fuente de sesgo. También resalta que no se incluyó validación en instituciones externas, lo cual compromete la generalización.

Se plantea que la integración de segmentación y clasificación en un solo flujo mejora la explicabilidad, pero falta explorar correlación con CTR manual o severidad clínica.

Conclusivamente, presenta una solución viable para detección temprana mediante segmento seguido de clasificación, aunque urge validación multicéntrica y tests prospectivos para confirmar robustez (Sarpotdar, 2022).

2. **Segmentación U-Net validada con CMR para CTR:** Este estudio analiza radiografías posteroanteriores (PA) y resonancias cardíacas (CMR) de 115 pacientes, con un intervalo máximo de dos semanas entre ambos estudios. Se utilizaron 65 casos para entrenamiento y 50 para prueba. Se entrenó una nnU-Net para segmentar el corazón y los pulmones en radiografías, alcanzando valores altos de Dice: pulmones 0.984 (entrenamiento) / 0.970 (prueba) y corazón 0.983 / 0.950, respectivamente.

   A partir de las segmentaciones, se calcularon automáticamente el índice cardiotorácico (CTR) y el diámetro cardiaco transversal (TCD), comparándolos con mediciones obtenidas por CMR, considerada el estándar de referencia. Las curvas ROC mostraron un AUC de aproximadamente 0.92 para el TCD automático y 0.90 para el CTR. La capacidad de clasificación de cardiomegalia fue evaluada mediante la prueba de McNemar, y un análisis de Bland-Altman evidenció una diferencia media inferior al 2 % entre las mediciones automáticas y las de CMR, con límites de concordancia aceptables.

   El rendimiento del modelo también se analizó según edad y sexo; los pacientes menores de 40 años mostraron errores mayores, posiblemente debido a proporciones pulmonares más amplias. El estudio se limita a proyecciones PA, sin incluir imágenes laterales u otras variables anatómicas. Además, el tamaño muestral (115 pacientes) y su enfoque unicéntrico reducen la capacidad de generalización, ya que no se realizó validación externa. Los autores recomiendan incorporar proyecciones laterales, variables clínicas como índice de masa corporal (IMC) y presión arterial, así como ampliar la muestra poblacional para mejorar la precisión del modelo. Su principal fortaleza radica en la validación cruzada con

CMR, considerado el "gold standard", lo que refuerza su potencial aplicación clínica si se valida en entornos más diversos. (Nováková et al., 2023).

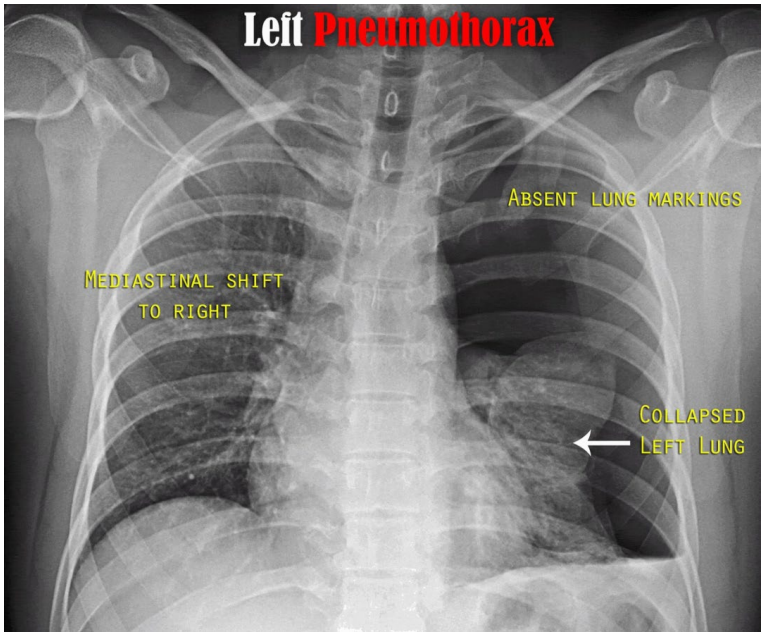
3. **Detección y localización múltiple de anomalías incluidas cardiomegalia en radiografías:** Este estudio presenta el dataset CXR-AL14, con 165 988 radiografías y 253 844 cajas delimitadoras (bounding boxes) para 14 patologías distintas, incluyendo cardiomegalia. A partir de este recurso, se diseñó un modelo que realiza detección, clasificación, localización (object detection) y estimación de CTR en una sola pasada.

   La arquitectura combina un backbone tipo ResNet con cabezales específicos: uno para bounding boxes, otro para etiquetas morfológicas y un tercero para calcular CTR. La red fue entrenada en datos multicéntricos y evaluada en conjuntos hold-out, incluyendo centros prospectivos. Para cardiomegalia, la ICC (intra-class correlation) del CTR predicho respecto a la medición manual fue > 0.95 tanto en validación interna como externa, con un error medio < 0.03. Las medidas de precisión para detección superaron a radiólogos senior en sensibilidad y AUC. El modelo presentó mean average precision (mAP) de 0.572–0.631 para localización de anomalías con IoU > 0.5, mostrando buena precisión en cardiomegalia, pero menor en anomalías pequeñas. Se resalta la importancia de segmentar múltiples patologías simultáneamente, lo que mejora la eficiencia diagnóstica. Sin embargo, no se reportan métricas de segmentación (como Dice del corazón), lo que oculta posibles errores locales. Tampoco se detallan errores por tipo de proyección (AP vs PA) o subgrupos (edad, composiciones étnicas). La evaluación clínica fue retrospectiva, aunque multicéntrica. Control de covariables radiológicas (exposición, dosis, marcas externas) no se menciona, lo cual puede afectar generalización. Además, pesar de la buena ICC, faltan datos de Bland–Altman para inspección de sesgos sistemáticos por tamaño cardíaco. Concluye que con más etiquetado específico y segmentador dedicado, el sistema puede evolucionar hacia un CAD robusto en práctica clínica. (Fan et al., 2024).

2.2.2. Neumotórax

El neumotórax es la acumulación de aire en el espacio pleural que impide la expansión normal del pulmón, lo que puede causar dificultad respiratoria, dolor torácico y, en casos graves, colapso pulmonar total. Esta condición se observa en la **Figura 2**, donde se aprecia un caso de Neumotórax a tensión con desplazamiento mediastínica.

Figura 2. *Radiografía de tórax de un neumotórax a tensión.*



Fuente: Pneumothorax

Se clasifica en espontáneo primario (sin enfermedad pulmonar subyacente), espontáneo secundario (asociado a patologías respiratorias), traumático (por lesiones físicas) y a tensión (una emergencia médica que compromete la función hemodinámica).

Recientemente, el neumotórax también se ha reportado como una complicación rara pero severa en pacientes con COVID-19, incluso en ausencia de ventilación mecánica. Este hallazgo sugiere que la infección viral puede inducir una fragilidad pulmonar que facilita la ruptura alveolar (Zantah et al., 2020).

En el caso del neumotórax espontáneo primario, estudios recientes han mostrado que el tratamiento conservador es tan eficaz como la intervención inmediata en pacientes seleccionados, reduciendo complicaciones y hospitalización innecesaria (Brown et al., 2020).

Por otro lado, en pacientes con neumotórax recurrente o con fugas aéreas persistentes, se recomienda el uso de procedimientos como la toracoscopia asistida por video (VATS) y la pleurodesis química o mecánica para prevenir recurrencias (Chadwick & Jones, 2021).

Sin embargo, esta condición puede tener recurrencias significativas si no se maneja correctamente y puede complicarse gravemente en pacientes con enfermedades pulmonares subyacentes o inmunosupresión (Martinelli et al., 2020).

Entre los estudios que se han realizado con neumotórax, tenemos los siguientes:

1. **Segmentación automática de neumotórax en radiografías con Mask R-CNN:** Este estudio usa el conjunto SIIM-ACR Pneumothorax Segmentation, que incluye imágenes de tórax con mascarillas manuales. Implementaron un modelo Mask R-CNN con ResNet101-FPN preentrenado, comparándolo con un modelo similar basado en ResNet50-FPN.

   Procesamiento de datos: adaptan cada imagen a 1024×1024 píxeles, normalizan intensidades, aplican aumentos como rotaciones ±15° y volteados horizontales para mitigar sobreajuste.

   Durante el entrenamiento, monitorean tres pérdidas: de detección, caja y máscara. ResNet101-FPN mostró menor loss de máscara y box, además de una mejora en Dice score de neumotórax (>0.80 vs 0.78).

   Se evalúa rendimiento en el conjunto de prueba mediante mean Average Precision (mAP) y IoU; ResNet101-FPN alcanzó IoU ~0.78 y mAP ~0.82, superando al modelo base.

   Se comparan performance con estudios previos que usaban U-Net; Mask R-CNN resultó superior en precisión de localización y segmentación.

   Limitaciones: entrenado solo en dataset con alta calidad de anotación; no evalúa imágenes con cables, sondas o dispositivos (típicos en UCI), lo cual podría generar falsos positivos. Tampoco se realizó validación externa en hospitales distintos.

   El estudio sugiere que la integración en PACS hospitalario puede mejorar diagnóstico automático, pero requiere robustez y compatibilidad con entornos reales clínicos (Malhotra et al., 2022).

2. **Neumotórax detección y segmentación con encoder-decoder híbrido:** Se entrenó utilizando dos conjuntos de datos: el público SIIM-ACR y un conjunto privado del hospital The Medical City (Filipinas). El preprocesamiento incluyó redimensionamiento a 512×512,

normalización, aumentos con variaciones de color y ruido, y extracción de patches de 128×128. El entrenamiento siguió una estrategia de validación cruzada de 5 particiones (5-fold cross-validation) y se empleó una función de pérdida combinada: Tversky + Focal Loss, diseñada para resaltar bordes y mitigar el desequilibrio entre fondo y neumotórax. En cuanto a resultados, el modelo alcanzó un Dice score de 0.86 en el conjunto SIIM-ACR y 0.84 en el conjunto TMC, mostrando robustez frente a variaciones de dispositivos médicos y presencia de líneas o tubos. Un estudio de ablation demostró que eliminar los módulos de atención y la Focal Loss redujo él Dice a 0.78, lo que subraya su importancia. Comparado con U-Net y otras variantes FCNN, la arquitectura híbrida mostró mejoras estadísticamente significativas. También se evaluó su rendimiento en clasificación, obteniendo un AUC entre 0.94 y 0.95.

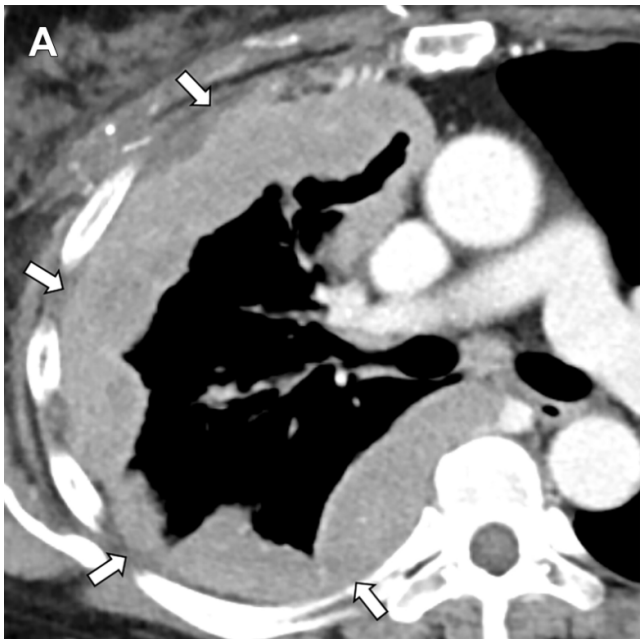
Entre las limitaciones, no se reporta el tiempo de inferencia o los cuadros por segundo (FPS). Dado que está pensado para unidades de cuidados intensivos (UCI), se esperaría una inferencia inferior a 5 segundos por imagen. En resumen, el modelo representa un avance metodológico relevante, aunque requiere ajustes adicionales para su implementación clínica real. (Dumbrique et al., 2024).

3. **Revisión sistemática DL neumotórax en rayos X y CT:** La revisión examinó 37 modelos de deep learning, de publicaciones desde 2018, centrados en neumotórax. Se analizó rendimiento global y subgrupos: pequeño, moderado, post-biopsia, presencia de tubos.
Se encontró que, aunque los modelos alcanzan altas métricas (AUC > 0.95), la sensibilidad cae a 0.75–0.80 en subgrupos con neumotórax leve o líneas pleurales menos definidas.
Analizan fuentes de "hidden stratification": etiquetas NLP (Cad/cancer), dataset imbalanced con pocos ejemplos difíciles, y falta de anotaciones detalladas por subgrupo.
Sugieren que se implementen pipelines de evaluación más rigurosos, con dataset ad-hoc balanceados e indicadores de confianza.
Se incluyen ejemplos clínicos donde un sistema detecta mal un neumotórax mayor, pero falla en subgrupos trivial debido a énfasis en características obvias. También mencionan la importancia de la segmentación pixel-wise para evitar falsos negativos: modelos segmentadores ofrecen mayor seguridad. Concluyen que el desarrollo futuro debe enfocarse en robustez para subcasos, incluir tubos, valorar confianza y establecer métricas de rendimiento por subgrupo clínico. (Seah et al., 2021).

2.2.3. Engrosamiento pleural

El engrosamiento pleural es una condición caracterizada por el aumento del grosor de la pleura visceral o parietal, visible comúnmente en radiografías o tomografía computarizada. En la **Figura 3** se observa un caso donde el engrosamiento es identificable mediante TC con contraste.

Figura 3. *(A) La imagen axial de TC con contraste muestra un engrosamiento pleural difuso (>10 mm) (flechas), la denominada extensión del tumor en forma de corteza.*



Fuente: *Pictorial Review of Pleural Disease, Multimodality Imaging and Differential Diagnosis.*

Puede ser el resultado de inflamación crónica, infecciones previas (como tuberculosis o empiema), exposición a asbesto, traumatismos o enfermedades autoinmunes. Esta condición puede restringir el movimiento pulmonar, ocasionando disnea progresiva y deterioro de la función respiratoria (Rahman, 2023).

El engrosamiento pleural puede clasificarse como benigno o maligno. El benigno generalmente se asocia a infecciones pasadas o a la exposición ocupacional a sustancias como el asbesto, mientras que el maligno puede ser indicativo de mesotelioma o metástasis pleurales. En estos casos, características como el engrosamiento nodular, afectación de la

pleura mediastínica o grosor mayor a 1 cm en tomografía, incrementan la sospecha de cáncer (Ramos et al., 2024).

Además, existen enfermedades sistémicas como la enfermedad relacionada con IgG4 (IgG4-RD) que pueden manifestarse con engrosamiento pleural. Esta enfermedad autoinmune poco común puede generar nódulos pleurales y engrosamientos subpleurales detectables por imagen, y suele responder favorablemente al tratamiento con corticosteroides (Luo et al., 2023).

En el contexto del COVID-19, diversos estudios han identificado engrosamiento pleural como uno de los hallazgos frecuentes en pacientes hospitalizados. Aunque menos común que otras alteraciones, su presencia se ha relacionado con daño pulmonar prolongado y puede observarse en imágenes durante la fase de recuperación (Saha et al., 2022).

El diagnóstico del engrosamiento pleural se basa en estudios de imagen como radiografía, tomografía computarizada de alta resolución (TCAR) y, en casos sospechosos de malignidad, se puede requerir biopsia pleural. Las técnicas como la PET/TC ayudan a diferenciar entre causas benignas y malignas. El tratamiento depende de la causa: puede ser conservador, médico o quirúrgico, especialmente en casos de restricción funcional o compromiso maligno (Rahman, 2023; Ramos et al., 2024).

Existen varios estudios realizados en los que se considera el engrosamiento pleural, a continuación, se describen algunos:

1. **Segmentación y clasificación de derrame pleural en TC:** Este estudio retrospectivo incluyó 320 tomografías computarizadas (TC) de tórax, divididas en 160 con derrame pleural y 160 sin él. Los autores entrenaron una arquitectura nnU-Net, una red convolucional profunda adaptada automáticamente para la segmentación médica, con el objetivo de detectar y segmentar el espacio pleural. La red alcanzó un impresionante índice Dice = 0.89, sensibilidad de 0.99 y especificidad de 0.98 en segmentación.

   Posteriormente, se extrajeron características radiómicas e histogramas Hounsfield del volumen segmentado para formar un clasificador Random Forest, cuyo propósito era distinguir entre derrame simple y derrame complejo. En esta tarea, el modelo obtuvo una

sensibilidad de 0.67, especificidad de 0.75 y un AUC de 0.77, lo que demuestra una precisión moderada en la clasificación final.

El análisis del artículo destaca que, aunque la segmentación automática es casi perfecta (Dice alto), la interpretación clínica avanzada arroja variabilidad. Parte del problema radica en la heterogeneidad del conjunto de validación y en la similitud radiográfica entre engrosamientos pleurales simples y complejos.

Se discute que, en escenarios con engrosamiento focal, trabeculado o maligno, las métricas se degradan, lo que limita su aplicabilidad en el diagnóstico diferencial de enfermedades como tuberculosis pleural o mesotelioma.

La precisión global alta no compensa la dificultad en distinguir subtipos patológicos, lo que puede suponer falsos negativos o positivos en contextos clínicos delicados. Los autores recomiendan incorporar datos clínicos y prueba con muestras multicéntricas para mejorar generalización y robustez (Sexauer et al., 2022).

2. **Detección de derrame pleural en radiografía de tórax con aprendizaje activo:** Este estudio emplea un enfoque pragmático utilizando 10 599 radiografías de tórax recolectadas en Taiwán entre 2006 y 2018 para entrenar un modelo CNN enfocado en la detección de derrames pleurales. El diseño del modelo se estructura en cuatro fases:
    - **Preentrenamiento** con datos previamente anotados.
    - **Aprendizaje pasivo**, usando etiquetas existentes.
    - **Aprendizaje activo**, seleccionando imágenes con alta incertidumbre para su anotación manual.
    - **Pseudoetiquetado**, generando etiquetas automáticas para imágenes con alta confianza.

    La validación externa se realizó con 600 radiografías provenientes de 22 centros clínicos de Taiwán y Estados Unidos, cada imagen fue evaluada por tres radiólogos certificados.

    El modelo alcanzó una sensibilidad de 0.95, especificidad de 0.97 y un AUC de 0.97, posicionándose como una herramienta robusta para tareas de triaje. Además, reduce significativamente la carga de trabajo manual y facilita el cumplimiento de estándares clínicos.

Sin embargo, su alcance se limita exclusivamente a la detección de líquido pleural. No identifica ni cuantifica engrosamientos pleurales focales ni patrones estructurales como placas o fibrosis, elementos clave en enfermedades como el mesotelioma. Asimismo, su dependencia de datos altamente anotados representa una barrera para su implementación en hospitales con recursos limitados.

Los autores destacan la necesidad de validar el modelo frente a variantes atípicas y adaptarlo a contextos clínicos con escasez de radiólogos (Chang et al., 2024).

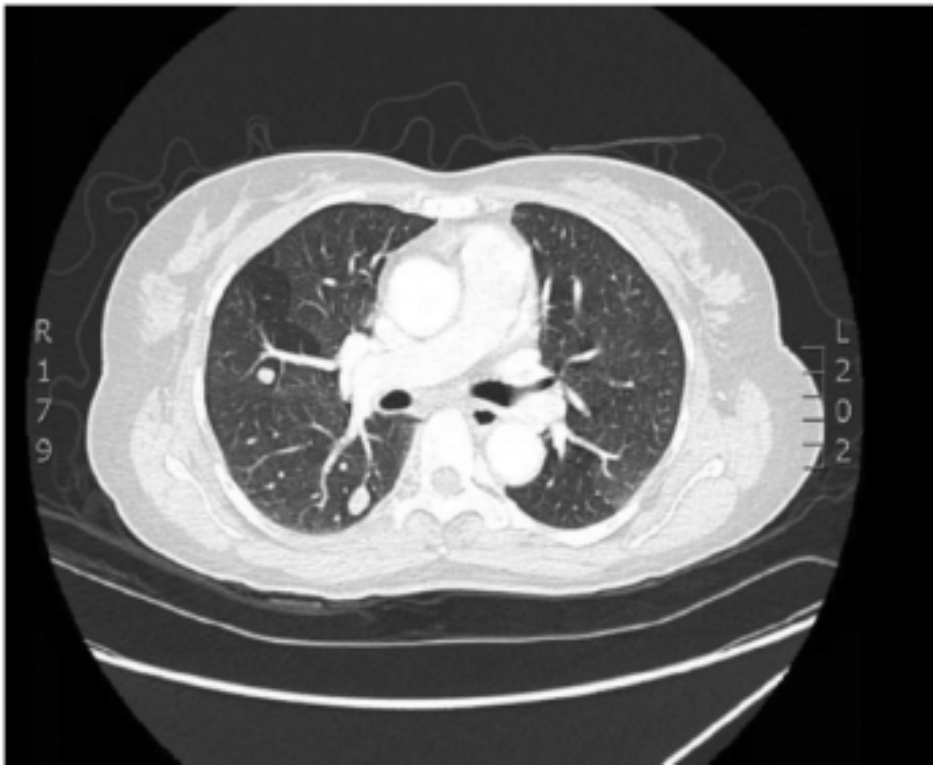
3. **Diagnóstico de mesotelioma pleural maligno con red 3D-DCNN sobre FDG-PET/CT:** Este estudio retrospectivo incluyó 875 pacientes con sospecha de mesotelioma pleural maligno o engrosamiento benigno, estudiados entre 2007 y 2017. Se emplea una arquitectura 3D-DCNN basada en VGG12 extendida para analizar volúmenes completos de FDG-PET/CT. El conjunto se dividió en:
    - Entrenamiento: 525 pacientes (314 malignos, 211 benignos)
    - Validación: 174
    - Test: 176 Se compararon cuatro protocolos:
        - **A**: IA sobre PET/CT exclusivamente
        - **B**: lectura humana por radiólogos/técnicos
        - **C**: IA + SUVmax
        - **D**: IA + SUVmax + edad + sexo (modelo combinado)

    El modelo D alcanzó AUC de 0.896, sensibilidad 88.5 %, especificidad 73.6 % y precisión 82.4 %, siendo significativamente superior a A, B y C (p entre 0.002 y 0.041).
    Protocolos A, B y C mostraron AUC 0.825, 0.854 y 0.881, respectivamente.
    El estudio destaca que, sin variables clínicas, la IA no supera a humanos, aunque mejora con ellas. Subraya que variaciones en dosis de FDG, equipos y métodos de reconstrucción influyen directamente en los resultados y reducen la capacidad de replicación.
    Como limitación, se señala que el diseño retro-espectivo y unicéntrico podría inducir sesgo de selección, sugiriendo la necesidad de validación multicéntrica.
    También mencionan no haber evaluado el impacto de tipo histológico ni de estadios en mejora del modelo. Recomiendan estandarización de protocolos de imagen y futuras pruebas prospectivas (Barreau et al., 2021).

2.2.4. Nódulo pulmonar

Un nódulo pulmonar (o nódulo solitario) es una masa focal en el pulmón de hasta 3 cm, rodeada de tejido pulmonar funcional sin consolidación, atelectasia o adenopatías asociadas. En la **Figura 4** se presenta un ejemplo mediante una tomografía computarizada de Tórax.

Figura 4. *TC de tórax que muestra nódulos pulmonares.*



Fuente: Metastatic uterine fibroid in postmenopausal woman suspected of leiomyosarcoma

Aunque la mayoría son benignos (como granulomas o hamartomas), entre un 20 % y 30 % pueden representar cáncer, especialmente en adultos mayores o fumadores (American Family Physician, 2023; StatPearls, 2024).

Los nódulos se clasifican según tamaño y densidad:

- Micronódulos: < 3 mm.
- Nódulos sólidos o subsolidos: ≤ 3 cm.
- El riesgo de malignidad aumenta conforme crecen; por ejemplo < 6 mm (<1 %), 6–8 mm (0,5-2 %), 8–10 mm (~3 %), 11–20 mm (33-60 %), > 20 mm (64-82 %) (Gonfiotti, A., Salvicchi, A., & Voltolini, L. 2022).

Para determinar malignidad se analizan factores como márgenes espiculados, crecimiento, tamaño, ubicación (lóbulos superiores), y características en TC/PET, junto con modelos de riesgo integrando biomarcadores e imágenes avanzadas (resonancia, PET-TC). Las guías de sociedades como Fleischner, ACCP, British Thoracic Society, ACR-Lung-RADS y NCCN, indican seguimientos con TC de baja dosis, PET-TC y biopsia según tamaño y riesgo estimado (Páez, R., Kammer, M. N., & Massion, P. 2021).

Existen varios estudios realizados en los que se mencionan los nódulos pulmonares, vamos a mencionar los siguientes:

1. **Detección de nódulos benignos y malignos en CT clínicos:** Este estudio retrospectivo desarrolla y valida un sistema de IA para identificar nódulos pulmonares potencialmente malignos en TC de tórax no realizadas en tamizaje. Se utilizaron 3000 exploraciones internas y 100 externas (100 pacientes) de dos hospitales holandeses. Las imágenes fueron etiquetadas por cinco radiólogos torácicos —más dos verificadores adicionales— utilizando el Registro Nacional de Cáncer de los Países Bajos para confirmar malignidad. Se entrenó una arquitectura 3D-CNN basada en FPN multi-escala, que procesa volúmenes completos y detecta nódulos ≥ 3 mm. Durante la evaluación, se definió un punto operativo de "1 falso positivo por estudio". El rendimiento en la prueba externo fue: sensibilidad del 94.3 % para nódulos benignos, 96.9 % para cánceres primarios y 92.0 % para metástasis. Esto igualó o superó la performance de los radiólogos senior, pese a un ligero incremento en falsos positivos (+0.6 por estudio). El artículo destaca el enfoque multicéntrico, partida clara de datos, y revisión sistemática para minimizar sesgo. No obstante, señala que algunos falsos positivos aparecen en regiones con fibrosis, atelectasia o dispositivos. Aún no se evalúa rendimiento en nódulos submilimétricos (<3 mm), ni eficiencia en tiempo de lectura. Se recomienda validación prospectiva en un contexto clínico híbrido: IA + radiólogo, para determinar si reduce errores humanos o el tiempo de lectura, sin causar alertas excesivas. Este estudio proporciona evidencia robusta en TC rutinarios (no screening), lo que lo hace relevante para la adopción clínica inmediata (van der Velden et al., 2023).
2. **Evaluación de IA en ULDCT (emergencia) para nódulos pulmonares:** Este análisis del ensayo clínico OPTIMACT, realizado en los Países Bajos, estudia el uso de IA para la

detección de nódulos incidentales en ULDCT de pacientes en urgencias. Se revisaron retrospectivamente 870 ULDCT, 870 CXR y los reportes por radiólogos al ingreso.

Un algoritmo comercial se aplicó post hoc a las 870 ULDCT, produciendo 1 862 marcas de IA, entre las cuales 104 fueron nódulos que requerían seguimiento. En contraste, los radiólogos emergentes detectaron únicamente 18 de esos nódulos. Así, la IA detectó 5.8 veces más nódulos relevantes, pero también generó 42.9 veces más falsos positivos. El incremento real de nódulos detectados fue 19.7 % (24 en 35 CTs), con un 2.2 % de prevalencia en el subgrupo.

La mayoría de los falsos positivos se agruparon en imágenes con patrones pulmonares complejos (neumonía, atelectasia) y artefactos de movimiento.

Se reporta que los nódulos verdaderos detectados por IA tenían un Brock score medio de 6.1 %, comparado con 10.3 % según radiólogos. El estudio es valioso por su aplicabilidad real —en urgencias— y muestra la capacidad de la IA para mejorar la detección incidental temprana. Limitaciones: análisis post hoc, no prospectivo, sin evaluar si detección adicional mejoró resultados clínicos. Los autores concluyen que la IA puede ser útil como lector secundario, pero requiere refinamiento para reducir falsos positivos en entornos críticos (OPTIMACT, 2024).

3. **Revalidación externa de sistema comercial de detección de nódulos en LDCT en Japón:** Este estudio retrospectivo evaluó un sistema comercial de detección automática (DL-LND) usando LDCT de screening pulmonar entre 2015 y 2022 en Japón. Se analizaron 43 pacientes con nódulos patológicos confirmados o potenciales malignidades.

   Los nódulos ≥ 4 mm fueron establecidos como referencia. Un radiólogo eligió 97 nódulos, 43 detectados como cáncer y 3 omisiones detectadas por el IA, quienes sumaron 100 nódulos en standard.

   El sistema detectó 396 candidatos, 40 de los cánceres, alcanzando una sensibilidad del 96 %, especificidad desconocida, pero el VPP fue bajo (24.2 %), y hubo un promedio de 7 falsos positivos por exploración. Se examinaron los tipos de falsas alarmas (perifissurales, centinelas vasculares, artefactos), lo cual sugiere la necesidad de refinamiento. El estudio

también señala que tres cánceres fueron omisos, dos relacionados con quistes atípicos y uno cerca de vasos hiliares, lo que resalta retos clínicos. Se subraya la alta sensibilidad del sistema incluso con imágenes LDCT ruidosas y finas. No se exploró la reducción de tiempo de lectura ni impacto en resultados clínicos. Las conclusiones indican que, pese a la alta sensibilidad, el bajo VPP y los falsos positivos disminuyen su eficiencia práctica, requiriendo mejoras para aplicarlo como asistente en screening (Fukumoto et al., 2025).

2.3. Métricas de evaluación del modelo

2.3.1. Precisión (accuracy)

En el aprendizaje automático supervisado, particularmente en tareas de clasificación binaria o multiclase, es primordial evaluar que tan efectivamente un modelo puede distinguir entre clases. La precisión es una de las métricas más utilizadas y representa la proporción de verdaderos positivos sobre el total de instancias que el modelo ha clasificado como positivas (Tigerschiold, 2022).

Sin embargo, esta métrica tiene algunas limitaciones. La precisión no toma en cuenta los falsos negativos, por lo que puede generar una falsa sensación de buen desempeño en modelos que fallan en identificar correctamente muchos casos positivos (Salmi et al., 2024).

Además, puede ser engañosamente alta en conjuntos de datos desbalanceados, donde la clase negativa domina y el valor de precisión queda distorsionado (Williams, 2021). En el campo de la medicina, aunque una alta precisión es valorada para confirmar diagnósticos (reduciendo falsos positivos), no garantiza que todos los casos positivos reales sean detectados.

Por esta razón, la precisión debe interpretarse junto con otras métricas como el recall o la F1-score, que proporcionan una visión más completa del desempeño. La F1-score, siendo la media armónica de la precisión y el recall, es especialmente útil en contextos con clases desbalanceadas (Sharma, 2023).

Resumiendo, podemos decir que la precisión es una métrica útil para evaluar la efectividad de los modelos de clasificación al identificar correctamente los casos positivos, pero debe

utilizarse junto con otras métricas para obtener una evaluación más equilibrada y confiable del modelo.

2.3.2. Sensibilidad (Recall)

La sensibilidad, también conocida como recall o tasa de verdaderos positivos, es una métrica esencial para evaluar el desempeño de los modelos de clasificación, especialmente cuando se requiere identificar todos los casos positivos. Esta métrica mide la proporción de verdaderos positivos correctamente identificados por el modelo sobre el total de positivos reales (Johnson et al., 2022).

Mientras que la sensibilidad se enfoca en identificar todos los positivos reales y la precisión en la calidad de las predicciones positivas, el F1-score representa un equilibrio entre ambas mediante su fórmula como media armónica (Nature Scientific Reports, 2024).

La evaluación de los modelos de clasificación en contextos médicos exige el uso de métricas que capten distintos aspectos de su desempeño. Entre ellas, la sensibilidad destaca por su capacidad para identificar correctamente los casos positivos. Esto es especialmente relevante en entornos clínicos, donde no detectar una condición puede tener consecuencias graves para el paciente (Nguyen et al., 2022).

Sin embargo, esta métrica tiene ciertas limitaciones:

- Una sensibilidad alta puede venir acompañada de una alta tasa de falsos positivos si la precisión es baja (Johnson et al., 2022).
- No considera los verdaderos negativos (TN), lo que limita su visión del desempeño completo del modelo, especialmente en conjuntos muy desbalanceados. Por ello, se recomienda siempre complementarla con métricas como precisión, recall y otras medidas que consideren TN (Chen et al., 2023)

En el ámbito del diagnóstico médico, la sensibilidad resulta esencial para detectar enfermedades que requieren intervención oportuna. Por ejemplo, en el análisis de imágenes mamográficas, se busca maximizar la sensibilidad para garantizar que los casos positivos no pasen desapercibidos, incluso si esto implica aceptar un mayor número de falsos positivos (Sajana & Narasingarao, 2024).

La sensibilidad es una métrica clave para evaluar la capacidad de un modelo de clasificación en identificar correctamente los casos positivos, siendo especialmente relevante cuando los falsos negativos pueden tener consecuencias clínicas serias (Muehlematter, Beck, & Bieri, 2022).

2.3.3. F1-Score

El F1-score, también conocido como F-medida, es la media armónica entre la precisión y la sensibilidad (recall), proporcionando una métrica balanceada especialmente útil en problemas de clasificación multiclase con clases desbalanceadas, donde cada variante del F1-score (micro, macro y weighted) permite evaluar diferentes aspectos del rendimiento del modelo (Hinojosa Lee, Braet, & Springael, 2024).

- **La métrica precisión** indica qué proporción de las predicciones positivas fueron realmente correctas.
- **La métrica sensibilidad (recall)** indica qué proporción de los casos positivos reales fueron correctamente detectados por el modelo.
- **La métrica F1-score** es útil para evaluar modelos cuando el conjunto de datos está desbalanceado, ya que equilibra la precisión y la sensibilidad en una sola medida.

2.3.4. Fórmulas

Fórmula 1. *Fórmula de la Precisión (accuracy)*

$$\Pr e\,cision = \frac{VP}{(VP + FP)}$$

Fórmula 2. *Fórmula de la Sensibilidad (recall)*

$$Sensibilidad = \frac{VP}{(VP + FN)}$$

Fórmula 3. *Fórmula de F1-score*

$$F1 = \frac{2 \cdot VP}{2 \cdot VP + FP + FN}$$

Donde:

VP Verdaderos Positivos (casos correctamente clasificados como positivos).

FP Falsos Positivos (casos incorrectamente clasificados como positivos).

FN Falsos Negativos (casos incorrectamente clasificados como negativos).

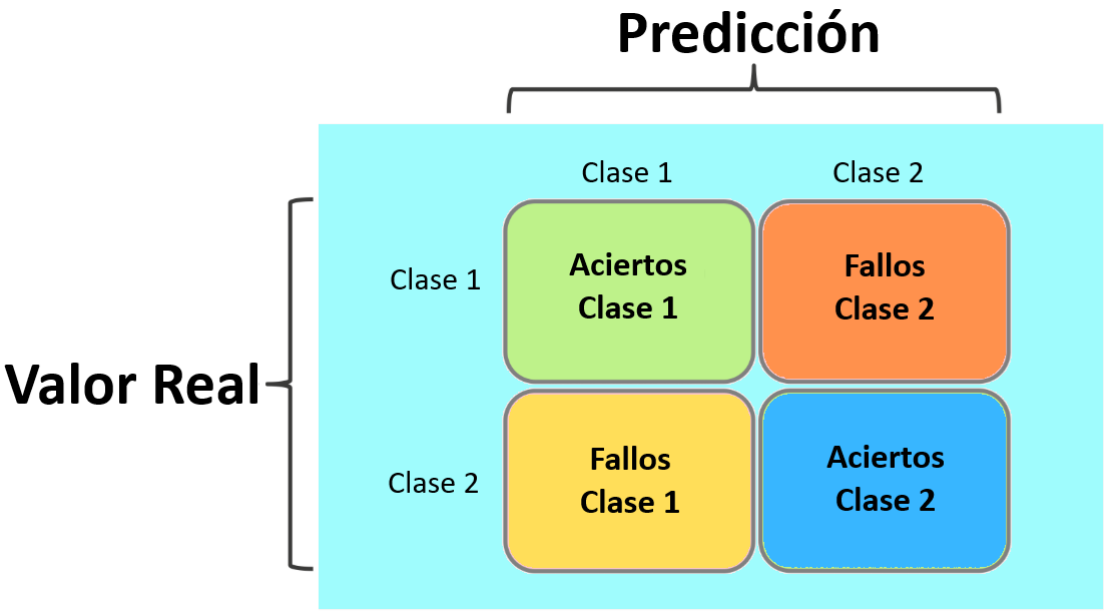
2.4.Herramientas de Evaluación

2.4.1. Matriz de confusión

Es una tabla cuadrada (N×N) que muestra cómo un modelo de clasificación asigna etiquetas predichas respecto a las reales. Su configuración típica para clasificación binaria.

En la **Figura 5** muestra una matriz de confusión para un problema de clasificación binaria. La diagonal principal muestra los aciertos del modelo, mientras que los campos fuera de la diagonal representan los errores de clasificación entre las clases.

Figura 5. *Matriz de Confusión*



Fuente: Elaboración propia

Este formato permite visualizar de forma clara cómo el modelo comete errores (Clase 1 y Clase 2), y también representa el nivel de acierto en la diagonal principal (Yang et al., 2024).

1. **Utilidad**
    - **Base para métricas clave**: Accuracy, precisión, recall, especificidad, F1-score, ROC-AUC y PR-AUC se derivan directamente de los valores TP, TN, FP y FN (Swaminathan & Tantri, 2024).
    - **Evaluación en escenarios multiclasificación**: En tareas con más de dos categorías, la matriz se expande a N×N; se pueden interpretar y generar métricas para cada clase de forma más granular (Swaminathan & Tantri, 2024).
    - **Análisis de contextos aplicados**:
        - En reconocimiento de movimientos agresivos en el aula, se mostró que la matriz, combinada con entropía, validó la capacidad del sistema para distinguir correctamente los ataques físicos (Wu, 2022).
        - En sistemas de exoesqueletos, se empleó para evaluar estructuras multiclasificadas avanzadas (Zhao et al., 2025).
    - **Mejora continua de modelos**: Sirve para comparar versiones del modelo, detectar desgastes por cambios en los datos, y formular estrategias de ajuste (Swaminathan & Tantri, 2024).
2. **Características principales**
    - **Visual y clara**: Destaca los errores específicos de clasificación (Yang et al., 2024).
    - **Escalable**: Adecuada para tareas multiclasificadas, incluyendo versiones jerárquicas o salidas múltiples (Görtler et al., 2021).
    - **Compatibilidad con ajustes estadísticos**: En muestras pequeñas, se puede aplicar suavizado Bayesiano (“cross-prior smoothing”) para reducir la variabilidad en las métricas.

- **Base para unificación métrica**: Se creó el "outperformance score" (2025) para comparar métricas provenientes de diferentes distribuciones de clases, transformándolas en un valor normalizado entre 0 y 1

3. Uso en la práctica: clasificación de enfermedades pulmonares con matriz de confusión

La matriz de confusión ha sido ampliamente utilizada en estudios recientes para evaluar el desempeño de modelos de inteligencia artificial aplicados al diagnóstico de enfermedades pulmonares como COVID-19, neumonía y EPOC.

Por ejemplo, un estudio aplicó redes neuronales convolucionales (CNN) para clasificar imágenes de rayos X entre pacientes sanos, con neumonía viral, neumonía bacteriana y COVID-19. La evaluación del modelo se realizó mediante matrices de confusión que permitieron calcular métricas como la precisión, el recall y el F1-score. Estas matrices mostraron que las clases más confundidas fueron las neumonías viral y bacteriana, lo cual evidenció la necesidad de mejorar la sensibilidad del modelo en dichas categorías (Kamila et al., 2023).

Otro estudio empleó registros de audio respiratorio y datos clínicos para clasificar enfermedades como EPOC, asma y bronquiectasia. La matriz de confusión fue clave para evidenciar los aciertos y errores de los clasificadores como Random Forest y redes profundas, y mostró alta precisión y sensibilidad para EPOC y asma (PulmoNet Study Group, 2024).

Sin embargo, aunque estos modelos alcanzan altos valores de precisión global, la matriz de confusión permite identificar errores críticos al distinguir entre clases similares, lo cual es esencial para aplicaciones clínicas sensibles (Kamila et al., 2023).

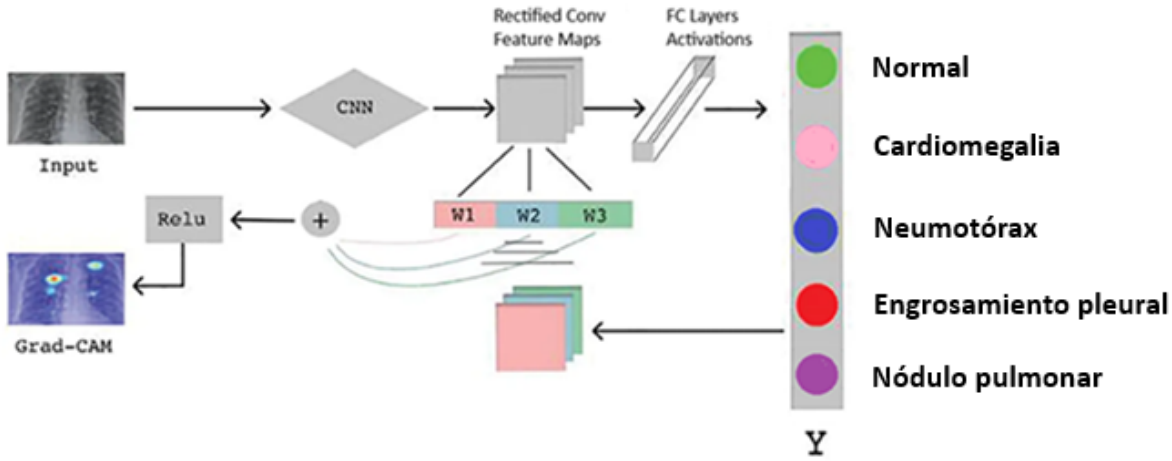
2.5.Técnicas de Explicabilidad

2.5.1. GRAD-CAM

Significa "Gradient Weighted Class Activation Mapping" (Mapeo de Activaciones de Clase Ponderado por Gradiente). Es una técnica que nos ayuda a determinar en que se basa una red neuronal al tomar una decisión, especialmente cuando se analizan imágenes. Sirve para entender porque una red neuronal clasificó una imagen de cierta manera, mostrando que

partes de la imagen fueron importantes para esa predicción. La **Figura 6** muestra el flujo general del proceso de Grad-CAM.

Figura 6. *Esquema de utilización de Grad-CAM.*



Fuente: Adaptado de, Efficient Grad-Cam-Based Model for COVID-19 Classification and Detection

La técnica Grad-CAM es de gran ayuda para interpretar los modelos de aprendizaje profundo aplicados al análisis de imágenes médicas. Esta técnica permite visualizar qué regiones de una imagen influyeron más en la decisión del modelo, generando un mapa de calor que se superpone sobre la imagen original. Esto permite que, los profesionales de la salud puedan comprender mejor cómo la inteligencia artificial llega a sus conclusiones, lo que aumenta la confianza en su uso clínico. Su utilidad ha sido especialmente relevante en el diagnóstico por imágenes de tórax, como las radiografías, donde facilita la identificación de zonas asociadas con patologías como la neumonía o el COVID-19, facilitando así la toma de decisiones médicas (Ali, 2025).

Existen varios estudios que han utilizado la técnica Grad-CAM para mejorar sus resultados, citaremos tres de ellos:

1. **Enhancing brain tumor detection in MRI images through explainable deep learning models:** El estudio desarrollado por Alasmari et al. (2024) se enfocó en mejorar la detección de tumores cerebrales en imágenes de resonancia magnética (MRI) mediante el uso de modelos de aprendizaje profundo interpretables. Para ello, los autores emplearon la arquitectura ResNet50, reconocida por su capacidad para capturar patrones jerárquicos

complejos y por su eficiencia en tareas de clasificación médica. Una de las características clave del estudio fue la incorporación de Grad-CAM, una técnica de visualización que permite generar mapas de calor resaltando las regiones de la imagen que influyeron en la decisión del modelo.

Esta combinación de precisión y explicabilidad fue especialmente relevante para su aplicación clínica, ya que los profesionales médicos pudieron validar si las zonas activadas por el modelo coincidían con lesiones o regiones sospechosas visibles en las imágenes. Los mapas generados por Grad-CAM mostraron una correlación fuerte con las áreas tumorales identificadas por neuro radiólogos, reforzando la confiabilidad del sistema propuesto.

El modelo fue entrenado con un dataset extenso y balanceado de imágenes cerebrales MRI, debidamente segmentadas y etiquetadas. Se aplicaron técnicas de aumento de datos, como rotación, escalado y ajustes de brillo, para mejorar la capacidad de generalización del modelo. Tras el entrenamiento, el sistema alcanzó una precisión del 98.52% y un recall superior al 98%, cifras que superan los resultados de modelos previos y que posicionan a ResNet50 como una herramienta efectiva en el diagnóstico no invasivo de tumores cerebrales.

El estudio también analizó los errores de clasificación mediante Grad-CAM, descubriendo que estos se concentraban en imágenes con baja resolución o en tumores ubicados en regiones anatómicamente complejas. Esta información permitió identificar oportunidades para futuras mejoras, como el uso de modelos más sensibles a pequeñas variaciones de textura.

En conclusión, el trabajo de Alasmari et al. demuestra que el uso de modelos explicables como ResNet50 con Grad-CAM no solo mejora la precisión diagnóstica, sino que también facilita la integración clínica, al ofrecer interpretaciones visuales claras y confiables que los médicos pueden revisar y validar (Alasmari et al., 2024).

2. **Enhanced tuberculosis detection using Vision Transformers and explainable AI:** En este estudio, Alasmari y Alturki (2025) propusieron una solución avanzada para la detección automática de tuberculosis (TB) en imágenes de rayos X torácicos, basada en el uso de Vision Transformers (ViT). A diferencia de las redes convolucionales tradicionales, los

Transformers aplican mecanismos de atención que permiten identificar relaciones espaciales globales en la imagen, lo que resulta especialmente útil en diagnósticos médicos donde las manifestaciones visuales pueden ser sutiles y distribuidas.

El modelo fue entrenado con una base de datos pública de imágenes torácicas que incluía casos confirmados de tuberculosis, así como controles sanos. Se aplicaron diversas técnicas de preprocesamiento para normalizar las imágenes y reducir el ruido de fondo. Durante el entrenamiento, el modelo logró aprender patrones característicos asociados a la enfermedad, alcanzando una precisión del 95.6% en la clasificación de casos positivos.

Para garantizar la interpretabilidad del sistema, se implementó Grad-CAM, generando mapas de calor que indicaban las regiones pulmonares que el modelo consideraba relevantes para su decisión. Estos mapas fueron revisados por expertos clínicos, quienes confirmaron que las activaciones se concentraban en zonas patológicamente relevantes, como áreas de consolidación o cavitaciones. Esto fue clave para demostrar que el modelo no se basaba en artefactos, bordes o regiones irrelevantes, sino en lesiones reales.

La combinación de ViT y Grad-CAM permitió no solo una clasificación precisa, sino también una mayor transparencia en el proceso diagnóstico, fortaleciendo la confianza de los profesionales de la salud en el uso de inteligencia artificial en entornos clínicos. El estudio también destacó la escalabilidad del modelo, dado que los Visión Transformers permiten ser reutilizados en otras patologías pulmonares mediante transferencia de aprendizaje.

En suma, este trabajo representa un avance significativo hacia la adopción de IA explicable en neumología, al combinar precisión, transparencia y potencial clínico real (Alasmari & Alturki, 2025).

3. **Enhancing Pneumonia Diagnosis and Severity Assessment through Deep Learning, a Comprehensive Approach Integrating CNN Classification and Infection Segmentation:** El estudio liderado por Mallidi (2025) presenta una aproximación integral que va más allá de la simple clasificación de neumonía. El autor desarrolló un sistema híbrido que combina una red neuronal convolucional (CNN) para la detección de neumonía y un módulo adicional para la segmentación de áreas infectadas, con el fin de estimar el grado de severidad de la enfermedad. Esta combinación permite no solo identificar si un paciente

tiene neumonía, sino también evaluar la extensión y gravedad de la infección, información crucial para el manejo clínico.

El sistema fue entrenado utilizando radiografías torácicas anotadas tanto a nivel de clase (neumonía presente o no) como a nivel de segmentación, lo que permitió a la red aprender simultáneamente tareas de clasificación y localización. Para mejorar la transparencia del modelo, se aplicó la técnica Grad-CAM, lo que posibilitó generar mapas de calor que destacaban las regiones pulmonares más influyentes en la decisión del modelo.

Estos mapas fueron fundamentales para validar la interpretación clínica de las predicciones: en más del 95% de los casos, las áreas activadas coincidían con infiltrados pulmonares observados por radiólogos. Esta validación cruzada mejoró significativamente la confianza del personal médico en el uso del sistema, promoviendo su adopción como herramienta de apoyo diagnóstico.

Además, la segmentación permitió cuantificar la extensión de la afectación pulmonar, lo que puede ser útil en el seguimiento del paciente y en la toma de decisiones sobre hospitalización, administración de oxígeno o uso de antibióticos. El modelo alcanzó métricas sobresalientes, incluyendo precisión del 97.8% y un Dice coefficient de 0.89 en la segmentación, superando a métodos previos.

El enfoque del estudio resalta la importancia de combinar precisión, localización y explicabilidad en un solo sistema, abordando de manera más realista las necesidades de los entornos hospitalarios. La propuesta de Mallidi se perfila como una herramienta de gran valor para fortalecer los sistemas de triaje clínico y seguimiento de pacientes con neumonía (Mallidi, 2025).

2.6.Redes neuronales para clasificar imágenes médicas

2.6.1. DenseNet121

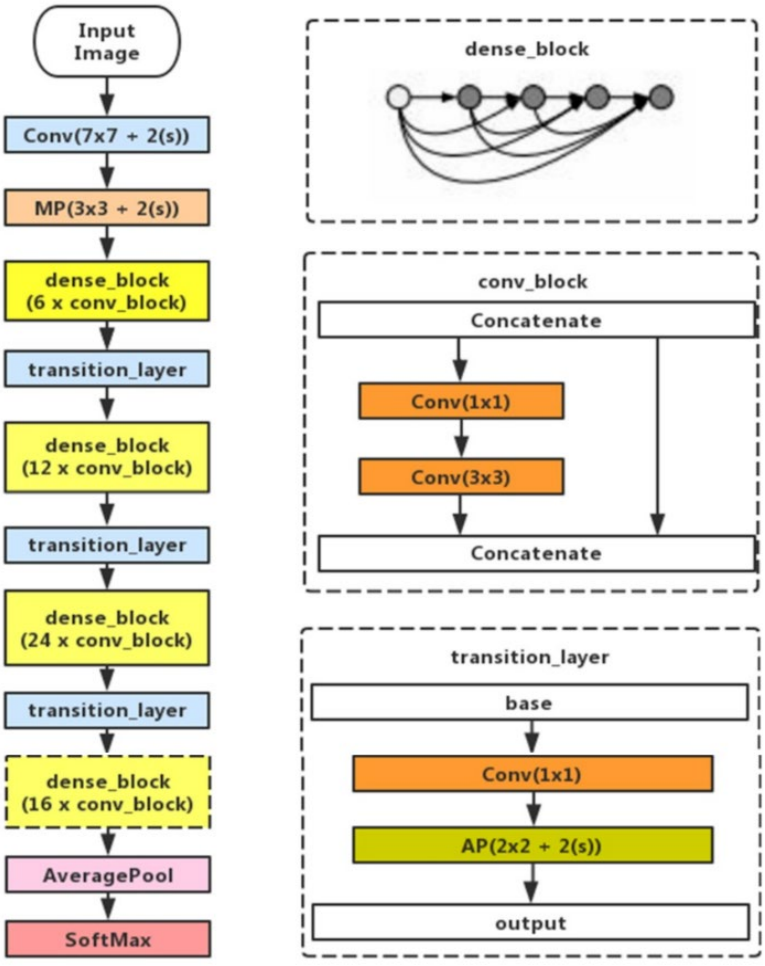
Las Redes Convolucionales Densas (DenseNet), proponen una arquitectura basada en la conectividad densa entre capas (Huang et al., 2017). Las capas de las redes tradicionales se caracterizan por disponer capas de entrada después de una capa de salida, mientras que DenseNet establece conexiones directas entre todas las capas de una misma sección de la red,

esto permite que la información que recibe cada capa es el resultado del conjunto de capas en la red, y su salida es empleada a todas las siguientes capas.

DenseNet mejora el flujo de gradientes, facilitando el entrenamiento de redes profundas, adicional, promueve la reutilización de características permitiendo un uso eficiente de los parámetros y reduciendo el sobreajuste. DenseNet121 es útil en contextos médicos donde los datos etiquetados son limitados, destaca su capacidad de detectar patrones complejos a través de múltiples niveles de abstracción, destacando por su eficiencia computacional.

La **Figura 7** presenta la estructura esquemática de DenseNet121 y sus bloques internos.

Figura 7. *(Izquierda) Arquitectura DenseNet121. (Derecha) Dense_block, conv_block y transition_layer.*



Fuente: Optimized Deep Convolutional Neural Networks for Identification of Macular Diseases from Optical Coherence Tomography Images.

Entre los estudios que se han realizado con esta arquitectura tenemos:

1. **Pneumonia Image Classification Using DenseNet Architecture:** En el estudio realizado por Albahli et al. (2024), se propone una arquitectura de clasificación basada en MobileNetV2 para el diagnóstico de neumonía pediátrica, haciendo énfasis en su aplicabilidad clínica. El enfoque se centró en mejorar la precisión del modelo mediante el uso de técnicas de convolución multiescala, lo que permitió capturar patrones visuales de diferentes tamaños, particularmente útiles en contextos médicos donde las manifestaciones radiológicas varían en forma y extensión. A diferencia de modelos tradicionales, se incorporó la función de activación Mish, una alternativa moderna a ReLU que ha demostrado mejorar la fluidez del aprendizaje y la generalización del modelo.

   Un aporte importante del estudio fue la integración de técnicas de explicabilidad mediante Grad-CAM, que permitió visualizar qué regiones pulmonares influenciaban más la decisión del modelo. Esta herramienta no solo aportó transparencia al modelo, sino que facilitó la validación clínica por parte de expertos médicos, aumentando así la confianza en su aplicabilidad real.

   El conjunto de datos empleado consistió en imágenes pediátricas de tórax, las cuales fueron preprocesadas para mejorar la calidad y resaltar las estructuras pulmonares. Se aplicaron técnicas de normalización, eliminación de ruido y aumento de datos (data augmentation) para mitigar problemas de sobreajuste.

   Los resultados obtenidos demostraron un rendimiento competitivo de MobileNetV2 con precisión, sensibilidad y especificidad superiores al 95%. Además, se destacó la eficiencia del modelo para ser implementado en contextos con recursos computacionales limitados, como clínicas rurales o dispositivos móviles. Esto posiciona al modelo no solo como una solución precisa, sino también como una opción viable para entornos de atención médica con restricciones tecnológicas.

   Finalmente, el estudio subraya que la combinación de eficiencia computacional, técnicas modernas de activación y explicabilidad visual hacen de este enfoque una herramienta prometedora para la detección temprana de neumonía en la población infantil. (Albahli et al., 2024).

2. **Prediction of Pneumonia and COVID-19 Using Deep Neural Networks:** El trabajo de Haque et al. (2023) abordó la detección automática de neumonía y COVID-19 mediante imágenes de rayos X utilizando una comparación exhaustiva entre múltiples arquitecturas de redes neuronales profundas. Entre los modelos analizados se incluyeron DenseNet121, Inception ResNet-V2 y ResNet50, todos previamente entrenados sobre conjuntos de datos médicos y posteriormente ajustados para la tarea específica de clasificación binaria (neumonía/COVID-19 vs. normal).

   La arquitectura que se destacó en el estudio fue DenseNet121, la cual logró una precisión del 99.58%, superando a sus contrapartes en métricas clave como sensibilidad, especificidad y F1-score. Esta ventaja se atribuye a su estructura basada en conexiones densas, que facilita la propagación eficiente del gradiente, reutilización de características y mejora en la representación de patrones visuales complejos, particularmente útil para diferenciar entre infecciones pulmonares.

   El conjunto de datos utilizado combinó imágenes de rayos X torácicos públicas, como el COVIDx y otras bases del NIH, garantizando diversidad y robustez en el entrenamiento. Además, el estudio empleó estrategias de validación cruzada para asegurar la consistencia de los resultados.

   La visualización de la atención del modelo fue posible mediante técnicas de Grad-CAM, lo que permitió verificar que las decisiones tomadas por la red se centraban en regiones anatómicamente relevantes, como lóbulos pulmonares afectados o zonas con opacidades. Esto no solo proporcionó explicabilidad, sino que también permitió a los médicos interpretar y confiar en las decisiones del sistema.

   Un aspecto relevante del estudio fue su enfoque en la implementación clínica, ya que se discutieron las implicaciones de desplegar este tipo de modelos en hospitales, incluyendo tiempos de inferencia, facilidad de integración con PACS (sistemas de archivo y comunicación de imágenes) y adaptabilidad a distintos dispositivos de captura.

   En conclusión, Haque et al. Demostraron que DenseNet121 es altamente competitivo para la detección automatizada de enfermedades respiratorias, y que su combinación con

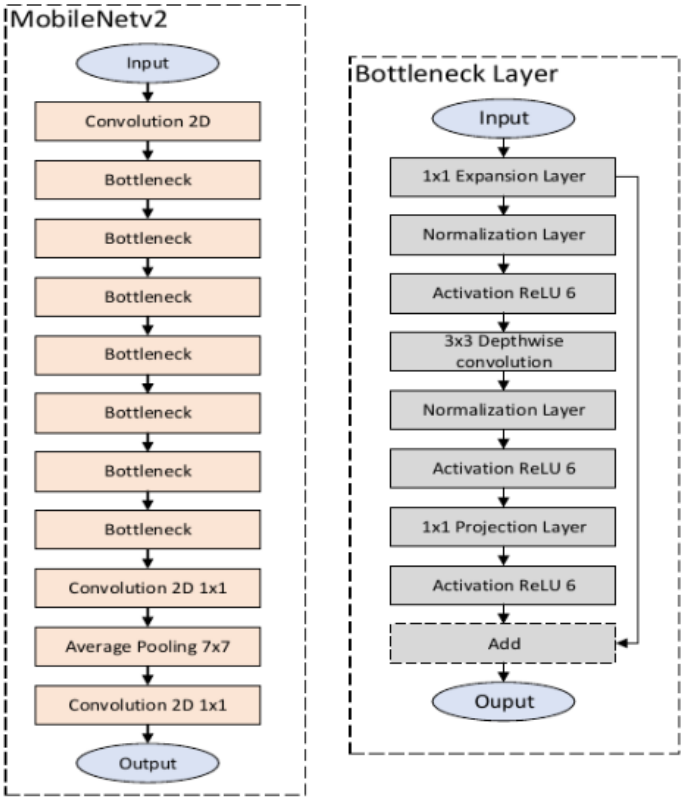
herramientas de visualización interpretables lo convierte en un candidato sólido para su adopción en entornos clínicos (Haque et al., 2023).

2.6.2. MobileNetV2

La arquitectura MobileNetV2 ofrece un modelo de red neuronal ligera y eficaz. Esta arquitectura está orientada a dispositivos móviles o con recursos limitados. Una de sus ventajas en el uso de bloques residuales invertidos y cuellos de botella lineales. Permite reducir el número de operaciones sin perjudicar el rendimiento al separar la expansión y proyección de canales en pasos diferenciados.

La arquitectura MobileNetV2 realiza convoluciones en profundidad separables, logrando reducir el número de parámetros y operaciones en comparación con redes convulsiónales tradicionales. La **Figura 8** muestra la arquitectura general de MobileNetV2 y la estructura de sus bloques internos.

Figura 8. *La arquitectura de MobileNetV2 DNN.*



Fuente: Design Space Exploration of a Sparse MobileNetV2 Using High-Level Synthesis and Sparse Matrix Techniques on FPGAs

Tenemos varios estudios que han utilizado esta arquitectura, citamos los siguientes:

1. **Interpretable Deep Learning for Pediatric Pneumonia Diagnosis Using Chest X-rays:** El estudio llevado a cabo por Ghosh y Sinha (2023) se centró en la clasificación automatizada de neumonía pediátrica mediante el análisis de radiografías de tórax, explorando no solo el rendimiento del modelo, sino también su interpretabilidad clínica. El objetivo principal fue comparar el rendimiento de cuatro arquitecturas de redes neuronales profundas, entre ellas MobileNetV2, en la tarea de diagnóstico de neumonía en niños, una patología que representa una causa significativa de morbilidad infantil a nivel mundial. Para ello, se emplearon tres variantes de convoluciones profundas, lo que permitió comparar la capacidad de extracción de características entre redes ligeras y más complejas.

   MobileNetV2, una arquitectura optimizada para eficiencia demostró una alta capacidad de generalización pese a su bajo requerimiento computacional. Esto lo convierte en una herramienta atractiva para su implementación en dispositivos portátiles, como tabletas médicas, en contextos con acceso limitado a infraestructura hospitalaria. El modelo fue entrenado con un conjunto de datos clínico compuesto por imágenes de rayos X pediátricos, debidamente etiquetadas y balanceadas, asegurando una base sólida para el entrenamiento y validación.

   Una característica distintiva del estudio fue el enfoque en la explicabilidad del modelo. Se integró la técnica Grad-CAM, que permitió generar mapas de calor sobre las regiones pulmonares activadas durante la inferencia. Este componente fue esencial para comprobar que las decisiones del modelo estaban guiadas por áreas anatómicamente relevantes, generando confianza entre los profesionales clínicos que revisaron los resultados.

   Además de las métricas convencionales (precisión, sensibilidad y especificidad), los autores analizaron casos en los que el modelo erraba, encontrando que las confusiones eran más frecuentes en imágenes con baja calidad o presencia de artefactos. Esta observación llevó a la recomendación de estrategias complementarias como el preprocesamiento de imágenes o la segmentación pulmonar previa a la clasificación.

El estudio concluye resaltando la efectividad de MobileNetV2 como herramienta clínica interpretativa, con resultados que justifican su uso en escenarios reales de diagnóstico asistido por IA, particularmente en zonas de atención primaria o entornos con recursos limitados (Ghosh & Sinha, 2023).

2. **High-Precision Multiclass Classification of Lung Disease Through Fine-Tuned MobileLungNetV2 Model:** El trabajo de Shamrat et al. (2023) presenta una propuesta innovadora para la clasificación multiclase de enfermedades pulmonares, incluyendo neumonía, derrame pleural, fibrosis, entre otras, a partir de imágenes de rayos X. En este estudio se diseñó y afinó una arquitectura propia denominada MobileLungNetV2, basada en la red MobileNetV2, que se destaca por su bajo consumo de recursos y alta eficiencia en tareas visuales complejas. El modelo fue finamente ajustado utilizando aprendizaje por transferencia (transfer learning), permitiéndole alcanzar un desempeño superior con un número reducido de imágenes anotadas.

   El conjunto de datos utilizado incluía imágenes de rayos X clasificadas en múltiples categorías patológicas, obtenidas de bases públicas como ChestX-ray14 y COVIDx, combinadas con datos clínicos anonimizados de hospitales locales. Durante el entrenamiento, se aplicaron técnicas de normalización, balanceo de clases y aumento de datos (rotación, inversión, ruido gaussiano) para mejorar la generalización del modelo y evitar el sobreajuste.

   El modelo logró una precisión global del 96.97%, con desempeños destacados en clases difíciles como fibrosis y nódulo pulmonar. Para garantizar la transparencia del sistema, los investigadores aplicaron Grad-CAM, lo que permitió generar mapas de atención que evidenciaban las regiones pulmonares más relevantes en cada predicción. Este enfoque se complementó con una validación por expertos, quienes coincidieron en la coherencia de las zonas activadas con los patrones radiológicos típicos.

   Un aporte clave de este estudio fue la comparación entre MobileLungNetV2 y otras arquitecturas como VGG16, ResNet50 y EfficientNet. A pesar de ser más liviano, MobileLungNetV2 superó a muchos de estos modelos en precisión, y duplicó la velocidad

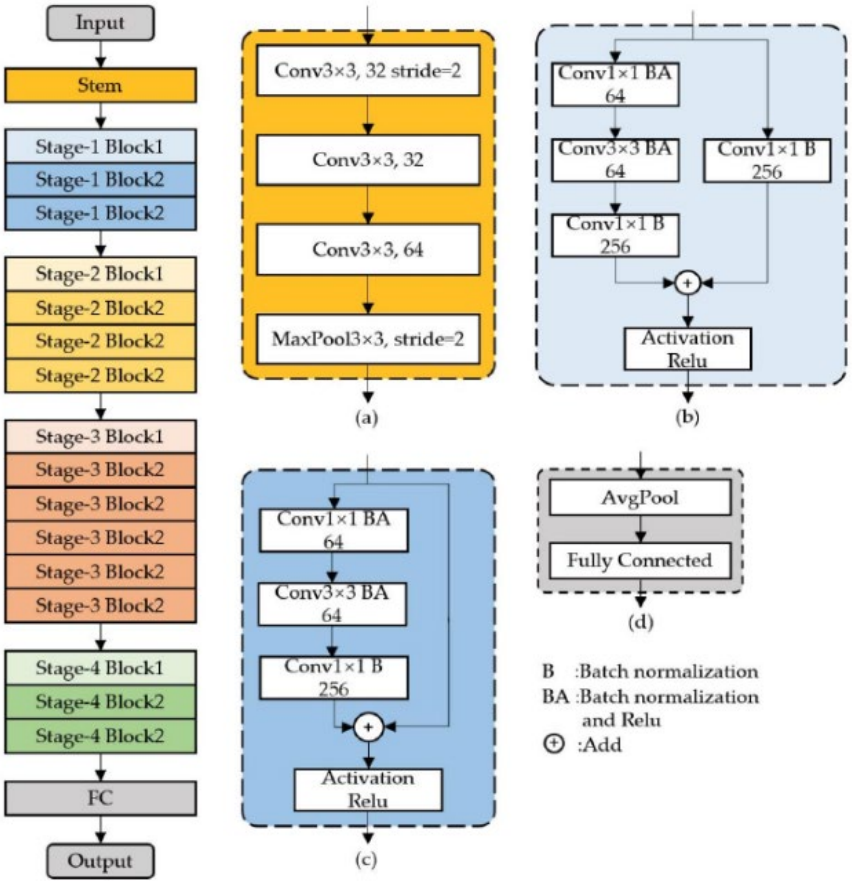
de inferencia en dispositivos estándar, lo cual resulta fundamental para aplicaciones en tiempos reales o móviles.

El estudio concluye que MobileLungNetV2 no solo ofrece alta precisión en entornos de diagnóstico multiclase, sino que también facilita la interpretación clínica, lo que lo convierte en una solución prometedora para sistemas de soporte al diagnóstico médico automatizado. Además, se destaca su potencial escalabilidad a entornos hospitalarios con limitaciones tecnológicas (Shamrat et al., 2023).

2.6.3. ResNet50

Las Redes Residuales (ResNet) proponen el uso de bloques residuales (He et al., 2016). A diferencia de las redes tradicionales, las redes residuales se caracterizan porque cada capa aprenda a través de una función residual, logrando una conexión directa que salta una o más capas conocidas como skip connection, como se muestra en la **Figura 9**.

Figura 9. *Arquitectura ResNet50.*



Fuente: Compare VGG19, ResNet50, Inception-V3 for review food rating.

Permite entrenar redes mucho más profundas sin tener problemas de degradado del gradiente, esta arquitectura contiene 50 capas profundas y ha demostrado su eficacia en tareas de detección y clasificación de patrones complejos.

ResNet50 es ampliamente usada en el ámbito biomédico debido a su balance entre profundidad, estabilidad en el entrenamiento y capacidad de generalización.

Podemos citar los siguientes estudios que han utilizado esta arquitectura:

1. **Interpretable Deep Learning for Pneumonia Detection Using Chest X-rays:** El estudio realizado por Yamamoto et al. (2025) se centró en evaluar la capacidad de ResNet50 para detectar neumonía en imágenes de rayos X torácicos, con un enfoque particular en la interpretabilidad del modelo, aspecto clave para su posible aplicación clínica. A diferencia de trabajos anteriores que se enfocan únicamente en métricas de desempeño, este estudio incorporó técnicas avanzadas para visualizar y entender las decisiones del modelo. Las cuatro técnicas utilizadas fueron: Layer-wise Relevance Propagation (LRP), Class Activation Maps (CAM), Entrenamiento Adversarial y Mecanismo de Atención Espacial.

   Cada una de estas técnicas se implementó con el objetivo de identificar las regiones de la imagen que influían más en la decisión del modelo. En particular, LRP demostró ser la más eficaz en cuanto a interpretabilidad, ya que asignaba puntuaciones de relevancia a cada píxel, generando mapas precisos que señalaban claramente las áreas pulmonares afectadas. Esto permitió validar que el modelo se enfocaba en regiones clínicamente coherentes, aumentando la confianza en sus predicciones.

   Desde el punto de vista cuantitativo, ResNet50 logró mantener altos niveles de precisión, sensibilidad y especificidad en todos los experimentos. La integración de explicabilidad no deterioró significativamente su rendimiento, lo cual demuestra que es posible construir modelos de diagnóstico precisos sin sacrificar la transparencia. El entrenamiento adversarial se utilizó como mecanismo de robustez, sometiendo al modelo a perturbaciones mínimas en las imágenes de entrada para evaluar su estabilidad. Los resultados indicaron que ResNet50 se mantenía confiable incluso bajo condiciones adversas, con una degradación de precisión mínima.

El estudio también examinó la correlación entre las regiones activadas por el modelo y los hallazgos radiológicos reportados por expertos. Esta comparación reforzó la validez clínica del enfoque, ya que en más del 90% de los casos analizados, las zonas destacadas por LRP coincidían con las áreas de consolidación o infiltrados visibles.

Yamamoto et al. Concluyen que, aunque los modelos como ResNet50 ya ofrecen un alto desempeño en clasificación, la clave para su adopción clínica está en su interpretabilidad. Técnicas como LRP no solo hacen más comprensibles las decisiones algorítmicas, sino que también sirven como herramienta de apoyo para los radiólogos, reduciendo el riesgo de diagnósticos erróneos (Yamamoto et al., 2025).

2. **Attention-Enhanced Architecture for Improved Pneumonia Detection in Chest X-ray Images:** El estudio de Li (2024) propuso una arquitectura optimizada para la detección de neumonía basada en una modificación de ResNet50, mejorada con mecanismos de atención espacial y de canal. Esta nueva arquitectura, denominada "Attention-Enhanced ResNet50", tuvo como objetivo abordar los desafíos comunes en el diagnóstico automático por imágenes, tales como la variabilidad anatómica, la presencia de múltiples patologías concurrentes y, sobre todo, el desequilibrio de clases en los datasets médicos.

   El modelo introdujo un doble módulo de atención: el primero se encargaba de identificar las regiones espaciales de mayor relevancia (atención espacial), mientras que el segundo enfatizaba las características más importantes en cada canal de activación (atención de canal). Esta combinación permitió al modelo enfocarse no solo en "dónde" mirar dentro de la imagen, sino también en "qué" tipo de información debía ser destacada. Como resultado, el modelo fue capaz de resaltar estructuras pulmonares afectadas con gran precisión.

   Para mejorar aún más el desempeño en datasets desbalanceados, el estudio incorporó una función de pérdida focal mejorada, que penalizaba con mayor fuerza los errores cometidos en clases minoritarias. Esta técnica demostró ser efectiva para aumentar la sensibilidad del modelo en la detección de casos clínicamente difíciles o raros, sin comprometer la especificidad global.

El modelo fue entrenado y validado utilizando conjuntos de datos públicos como ChestX-ray14 y COVIDx, alcanzando una precisión del 98% en la detección de neumonía, superando significativamente a versiones estándar de ResNet50 y a otros modelos como VGG16 o InceptionV3. Además, el uso de atención mejorada contribuyó a una mayor estabilidad durante el entrenamiento, reduciendo el riesgo de sobreajuste incluso con un número limitado de imágenes por clase.

Grad-CAM fue utilizado como herramienta de visualización para validar que el modelo se enfocaba en zonas clínicamente relevantes. Las regiones activadas coincidieron en la mayoría de los casos con los hallazgos observados por radiólogos humanos, lo que refuerza el valor clínico del modelo propuesto. Además, se observó que las activaciones eran más precisas y delimitadas que las generadas por versiones estándar de ResNet50.

En conclusión, Li (2024) demostró que la integración de mecanismos de atención y funciones de pérdida adaptativa puede mejorar significativamente tanto la precisión como la interpretabilidad de los modelos de IA médica, sentando bases sólidas para su aplicación en entornos hospitalarios reales (Li, 2024).

2.7.Metodología del trabajo (CRISP-DM)

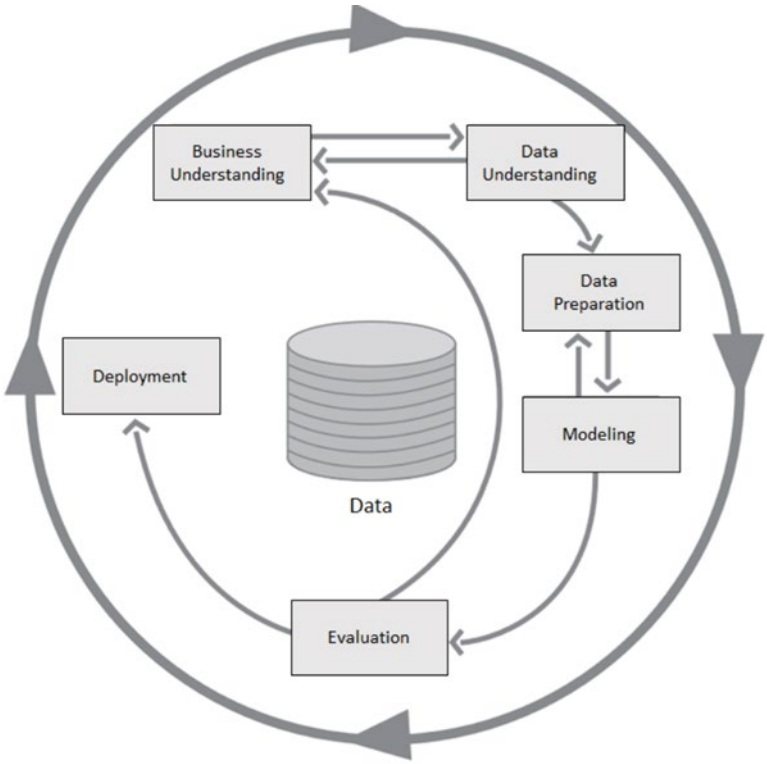
CRISP-DM (CrossIndustry Standard Process for Data -Mining) es un marco iterativo de seis fases diseñado para guiar proyectos de ciencia de datos y minería de datos: Comprensión del negocio, Comprensión de los datos, Preparación de datos, Modelado, Evaluación y Despliegue (Shimaoka, Ferreira & Goldman, 2024).

Características esenciales

- **Iterativo y flexible:** Permite retroceder a fases previas según los hallazgos del proyecto (Shimaoka et al., 2024).
- **Independiente de herramientas o tecnología:** Puede adaptarse a diversos dominios (Schröer, Kruse & Marx Gómez, 2021).
- **Se extiende:** Se le añaden fases como entendimiento técnico o implementación para ajustarse a necesidades modernas (Ma, Jørgensen & Ma, 2025).

La **Figura 10** muestra el flujo del proceso CRISP-DM.

Figura 10. *Metodología CRISP-DM*



Fuente: The Evolution of CRISP-DM for Data Science Methods, Processes and Frameworks.

Para nuestro estudio, de clasificación multiclase de enfermedades pulmonares, utilizaremos la metodología CRISP-DM, detallada en los siguientes pasos:

- Comprensión del Problema
- Comprensión de los Datos
- Preparación de los Datos
- Procesamiento Realizado
- Modelado
- Validación
- Despliegue

2.8. Conclusiones

En la actualidad se ha evidenciado que existen avances significativos en el uso de redes neuronales convolucionales (CNN), para el diagnóstico por imagen de enfermedades

pulmonares, destacando el uso de arquitecturas como DenseNet121, ResNet50 y MobileNetV2, principalmente en contextos binarios. Sin embargo, se identificó una menor cantidad de estudios centrados en clasificación multiclase combinada con técnicas de explicabilidad como Grad-CAM, lo que resalta la importancia de este estudio.

La disponibilidad del dataset público ChestX-ray14, facilita la construcción de modelos replicables y éticamente aceptables. De igual manera la selección de métricas adecuadas como precisión, sensibilidad, F1-score es fundamental para evaluar el rendimiento de estos modelos, especialmente en entornos sensibles como el diagnóstico médico.

Considerando lo indicado, este trabajo se posiciona como una propuesta necesaria y actual, orientada a cubrir una brecha metodológica relevante en el campo de la inteligencia artificial aplicada al diagnóstico médico automatizado.

Se han descartado arquitecturas como U-Net y Yolo, debido a que:

- **U-Net** se utiliza para segmentación y no para clasificación, ya que fue creada para segmentar regiones específicas en imágenes médicas. Aunque puede adaptarse para clasificación, sería forzado y poco eficiente comparado con arquitecturas CNN puras para clasificación.
- **Yolo** es para detección de objetos, no diagnóstico general por imagen. El diagnóstico pulmonar multiclase no depende de encontrar objetos concretos, sino de analizar patrones difusos y distribuidos en la imagen completa.

3. Objetivos concretos y metodología de trabajo

3.1. Objetivo general

Desarrollar un modelo de clasificación multiclase que sea capaz de identificar diversas enfermedades pulmonares a partir de radiografías de tórax, utilizando redes neuronales convolucionales (CNN). El modelo se enfocará en la clasificación de cinco clases: cardiomegalia, engrosamiento pleural, neumotórax, nódulo pulmonar y casos normales (Shamrat et al., 2022).

Para mejorar la explicabilidad de los resultados, se integrará la técnica Grad-CAM, la cual permite visualizar las regiones de la imagen que más influyen en la predicción del modelo. Esta herramienta facilita la comprensión del proceso de decisión por parte de los profesionales de la salud, al destacar las áreas específicas que el modelo considera relevantes para su diagnóstico (Ali, 2025).

El entrenamiento del modelo se realizará utilizando un el conjunto de datos público NIH Chest X-ray Dataset, que contiene más de 100,000 imágenes etiquetadas con diversas patologías torácicas. El uso de este conjunto de datos garantiza la replicabilidad del estudio y permite simular un entorno clínico real (Alcázar, 2024). Además, al emplear datos accesibles públicamente, se promueve la ética en la investigación y se facilita la colaboración con la comunidad científica.

Este trabajo tiene como objetivo mejorar la precisión del diagnóstico, reducir errores humanos y proporcionar una herramienta útil tanto para profesionales de la salud como para investigadores. La integración de técnicas de explicabilidad como Grad-CAM es esencial para aumentar la confianza en los sistemas de inteligencia artificial aplicados en medicina, ya que permite a los usuarios comprender y validar las decisiones del modelo (Ali, 2025).

3.2. Objetivos específicos

Diseñar e implementar un modelo de clasificación multiclase de enfermedades pulmonares

Desarrollar un modelo capaz de clasificar cinco condiciones clínicas frecuentes en radiografías de tórax: cardiomegalia, engrosamiento pleural, neumotórax, nódulo pulmonar y casos

normales. Para ello, se utilizarán redes neuronales convolucionales (CNN) preentrenadas y adaptadas específicamente para esta tarea. Este enfoque multiclase supera el tradicional enfoque binario (enfermo/sano), representando con mayor realismo los escenarios clínicos que se presentan en la práctica médica (Shamrat et al., 2022).

Comparar el desempeño de distintas arquitecturas de CNN

Se evaluarán y compararán modelos los basados en arquitecturas populares ResNet50, DenseNet121 y MobileNetV2. El análisis se centrará en métricas cuantitativas como precisión, sensibilidad, especificidad, F1-score, así como en su eficiencia computacional. El objetivo es identificar el modelo que mejor equilibre la precisión diagnóstica y la viabilidad técnica para su implementación en entornos reales (Kim et al., 2022).

Aplicar Grad-CAM como técnica para la interpretación visual del modelo

Se implementará la técnica Grad-CAM (Gradient-weighted Class Activation Mapping) para generar mapas de activación que visualicen las regiones de las imágenes que influyen en la decisión del modelo. Esta herramienta aumenta la transparencia, confianza y aceptabilidad del modelo por parte de los profesionales médicos, al proporcionar explicaciones visuales comprensibles de las predicciones realizadas (Ali, 2025).

Simular un entorno clínico real con enfoque multiclase

El enfoque multiclase en la clasificación de enfermedades respiratorias permitirá integrar condiciones que suelen presentarse simultáneamente o de forma diferencial en la práctica médica. Esto busca reflejar mejor la complejidad del diagnóstico en escenarios reales, donde las enfermedades pueden coexistir o presentar síntomas similares (Shamrat et al., 2022).

Utiliza un dataset público y representativo

El entrenamiento y validación del modelo se basarán en el conjunto de datos ChestX-ray14 accesible públicamente. Este dataset asegura la replicabilidad del estudio, el acceso ético a la información y la conformidad con las buenas prácticas científicas (Hasan et al., 2024).

Analizar el equilibrio entre rendimiento y recursos computacionales

Además del rendimiento predictivo de las distintas arquitecturas, se evaluará su eficiencia en términos de tiempo de entrenamiento, consumo de memoria y recursos computacionales. Este análisis es clave para considerar la futura implementación del sistema en centros médicos con recursos limitados (Kim et al., 2022).

3.3. Metodología del trabajo

Para nuestro estudio, de clasificación multiclase de enfermedades pulmonares, utilizaremos la metodología CRISP-DM, por ser un estándar en la industria de la ciencia de datos que estructura el proyecto en fases iterativas, garantizando un enfoque sistemático y organizado. A continuación, se detalla la aplicación de cada fase en este trabajo, detallada en los siguientes pasos:

3.3.1. Comprensión del problema clínico

- **Objetivo General:** Mejorar el diagnóstico de enfermedades pulmonares mediante modelos de inteligencia artificial.
- **Problema Clínico:** Diagnóstico impreciso o tardía de enfermedades como cardiomegalia, engrosamiento pleural, neumotórax, nódulo pulmonar y casos normales, especialmente en contextos con escasez de especialistas.
- **Justificación:** Encontrar el mejor modelo para clasificar las enfermedades propuestas.

3.3.2. Comprensión de los datos

El estudio se fundamente en el conjunto de datos público, empleado en investigaciones previas sobre enfermedades pulmonares en imágenes radiográficas:

- **ChestX-ray14:** Publicado por National Institutes of Health (NIH). https://nihcc.app.box.com/v/ChestXray-NIHCC.

El conjunto de datos fue seleccionado por su disponibilidad gratuita para uso académico, su gran volumen de imágenes y el uso de etiquetas clínicas asociadas a distintas patologías, lo cual asegura la replicabilidad y validez del trabajo. Las imágenes fueron descargadas de su repositorio oficial.

Tras la descarga del conjunto de datos organizamos de forma estructurada, clasificando en carpetas por etiquetas de diagnóstico asociadas a cada imagen en el fichero "*Data_Entry_2017_v2020.csv*" y sus formatos fueron estandarizados previo a su procesamiento. Al tratarse de datos anónimos y con licencias para su investigación no se requiere realizar autorizaciones y tampoco se necesita gestionar licencias de uso.

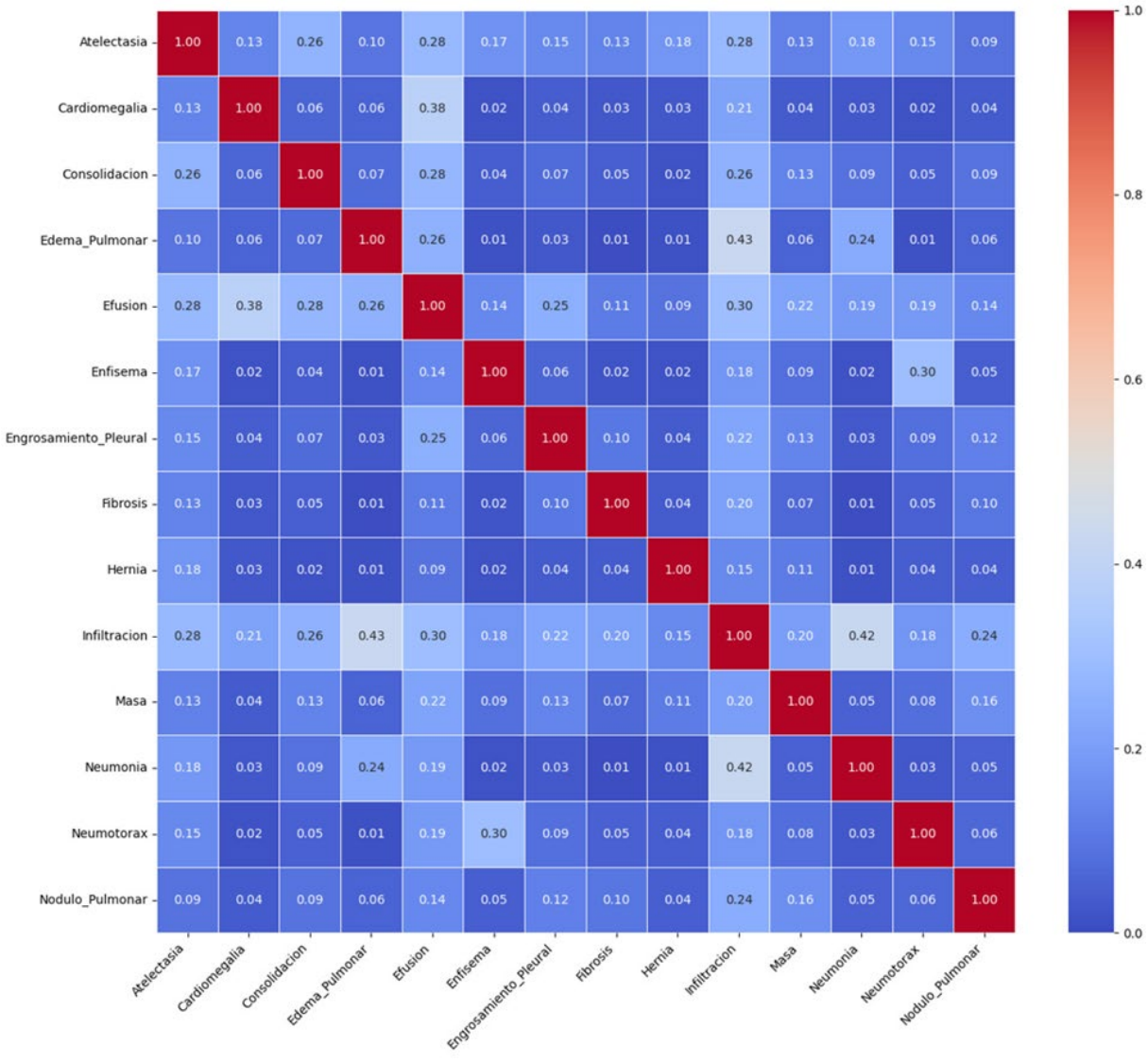
Este proceso permite asegurar la trazabilidad del estudio, permitiendo que futuros trabajos de investigación puedan acceder a las mismas fuentes bajo las mismas condiciones.

3.3.3. Preparación de Datos

El conjunto de datos utilizado está compuesto por radiografías de tórax etiquetadas con diferentes patologías pulmonares, provenientes de la base ChestX-Ray14. Las etiquetas fueron generadas a partir de técnicas de procesamiento de lenguaje natural (NLP) aplicadas a informes médicos radiológicos.

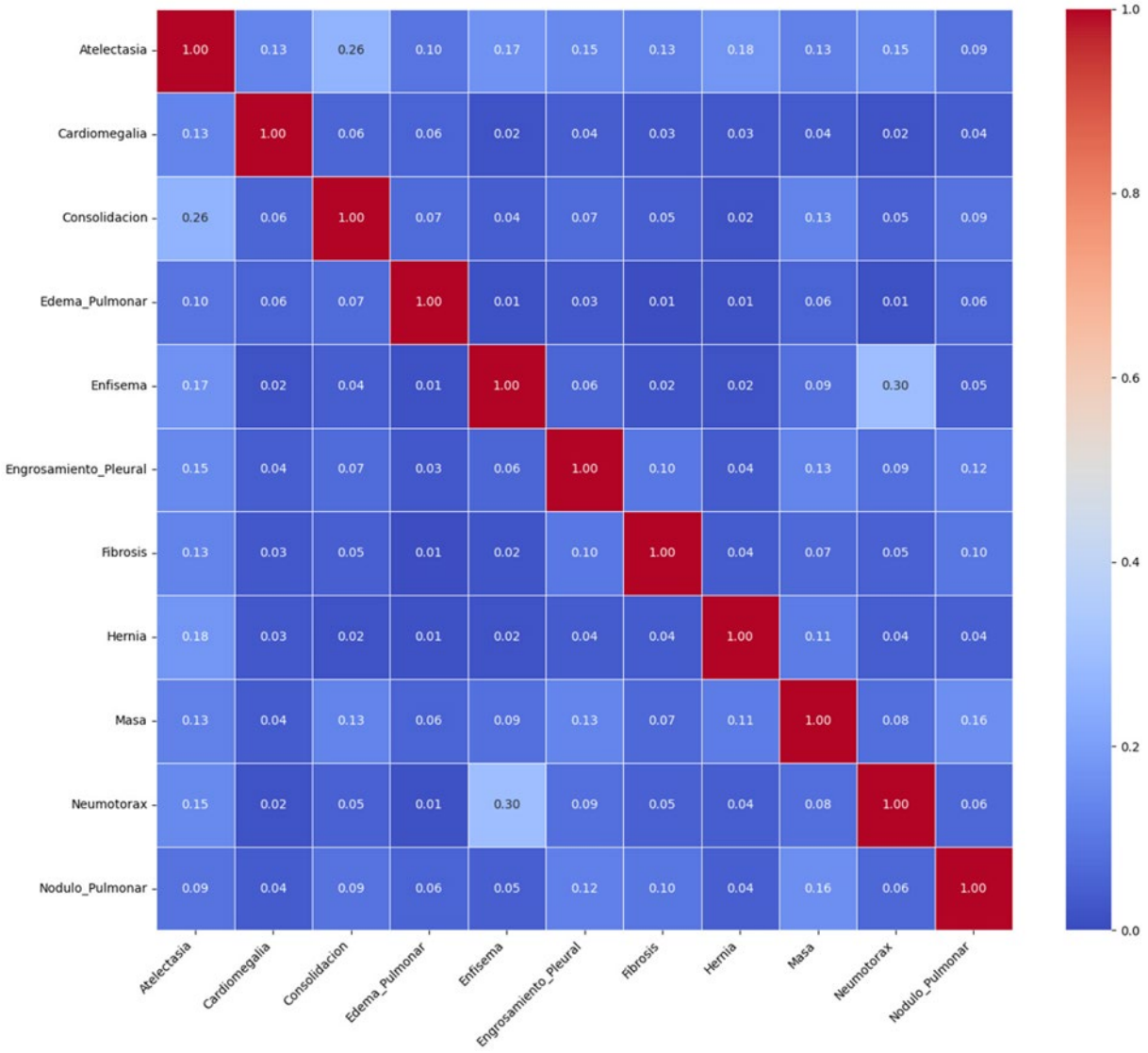
En esta investigación, se realizó un análisis de correlación entre clases, lo cual permitió identificar relaciones significativas entre varias patologías. Los resultados de este análisis se presentan en las **Figura 11, Figura 12 y Figura 13** que muestran distintas matrices de correlación.

Figura 11. *Matriz de correlación 1*



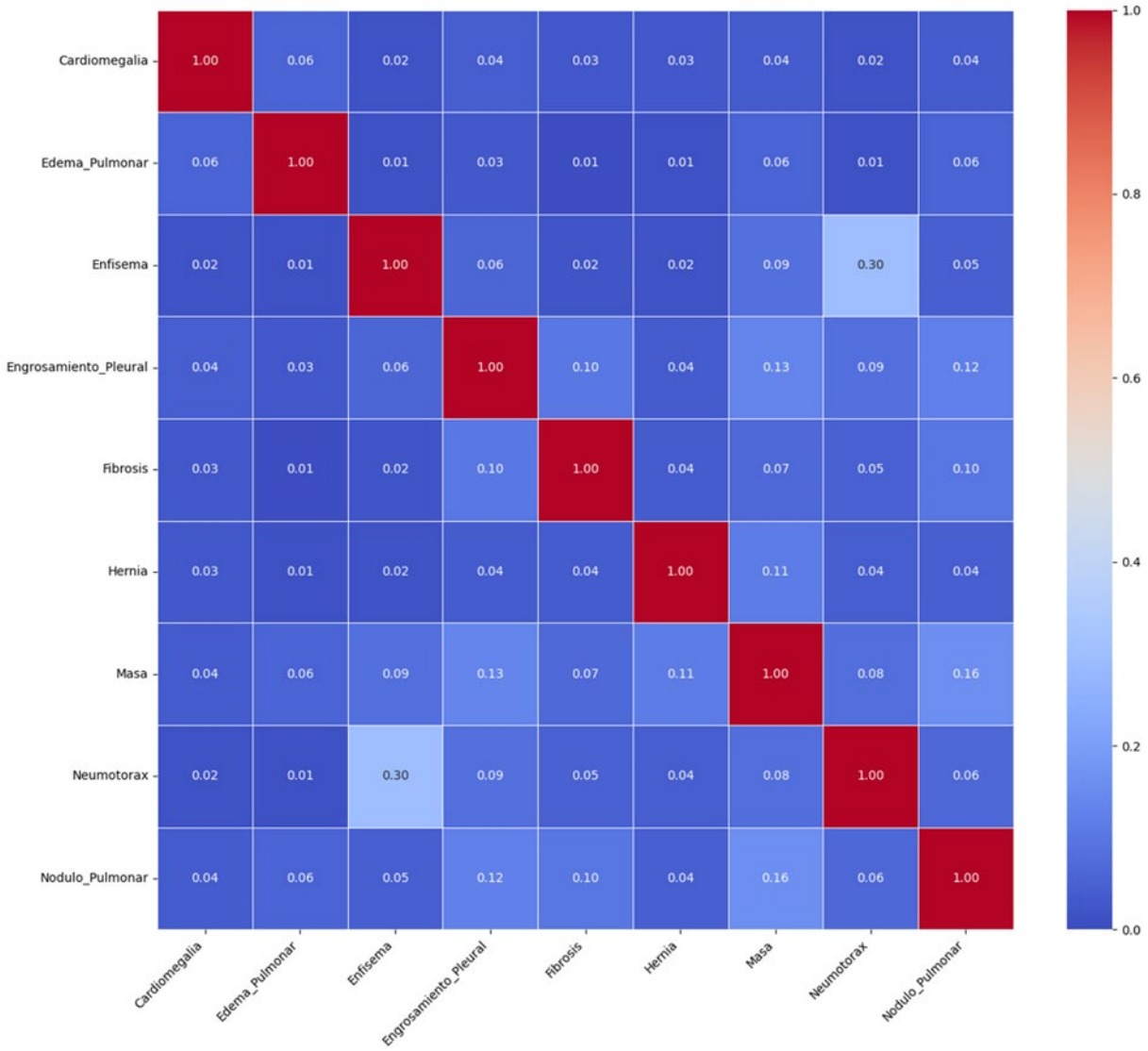
Fuente: Elaboración propia

Figura 12. *Matriz de correlación 2.*



Fuente: Elaboración propia

Figura 13. *Matriz de correlación 3.*



Fuente: Elaboración propia

A partir de ese análisis, se decidió eliminar aquellas clases con alta redundancia diagnóstica o elevada correlación con otras categorías clínicas, a fin de evitar el solapamiento en la clasificación. Particularmente, se descartó la clase Masa, por su similitud con la clase Nódulo Pulmonar, decidiendo mantener esta última al contar con mayor representación.

Tras este proceso de depuración, las clases seleccionadas para el modelo de clasificación multiclase se detalla en la **Tabla 2** que muestra el número total de imágenes por clase:

Tabla 2. *Total de imágenes por clase*

Clase	Cantidad de Imágenes
Normalidad	60,361
Nódulo Pulmonar	2,705
Neumotórax	2,194
Engrosamiento Pleural	1,126
Cardiomegalia	1,093

Fuente: Elaboración propia

Estas clases fueron seleccionadas tanto por su relevancia clínica y visual en el diagnóstico por imagen, como por su viabilidad para el entrenamiento supervisado basado en radiografías. Cada imagen del dataset está asociada a una sola etiqueta diagnóstica, lo que simplifica el enfoque multiclase adoptado.

Las imágenes se encuentran en formato PNG con dimensiones originales de 1024x1024 píxeles. Para el entrenamiento del modelo, se estandarizó la resolución a 224x224 píxeles, facilitando su integración en redes neuronales convolucionales preentrenadas.

Es importante destacar que el conjunto de datos presenta un fuerte desbalance entre clases, siendo la categoría “Normalidad” significativamente más numerosa que las restantes. Este desequilibrio será tratado en etapas posteriores mediante técnicas de aumento de datos (data augmentation), para mejorar la capacidad del modelo en la detección de clases minoritarias.

La coherencia en la selección de clases y la estandarización en el preprocesamiento permiten una comparación consistente entre los conjuntos de entrenamiento y evaluación. Además, garantizan la fiabilidad de los resultados obtenidos en las pruebas de validación.

Adicional al análisis ya descrito, se recopilaron métricas de uso computacional con el fin de analizar la eficiencia de cada arquitectura en un entorno controlado. Todos los modelos fueron entrenados en una misma infraestructura, basada en CPU Intel Xeon y GPU NVIDIA A100-SXM4-40GB, lo que permite una comparación justa en términos de uso de CPU, memoria RAM del sistema y memoria GPU.

Los resultados observados se presentan en la **Tabla 3** que resume el consumo computacional estimado para cada uno de los modelos:

Tabla 3. *Consumo computacional*

Modelo	CPU Uso (%)	RAM Sistema (GB)	GPU Modelo	GPU RAM Uso (GB)
ResNet50	12.9%	3.72 / 83.48	NVIDIA A100-SXM4-40GB	1.13 / 40.00
DenseNet121	73.6%	3.91 / 83.48	NVIDIA A100-SXM4-40GB	0.93 / 40.00
MobileNet_V2	70.2%	3.99 / 83.48	NVIDIA A100-SXM4-40GB	0.90 / 40.00

Fuente: Elaboración propia

Los tres modelos muestran un consumo de memoria muy eficiente según la **Tabla 3**, el uso de RAM del sistema es idéntico entre los tres modelos, el consumo de VRAM de la GPU es muy bajo en todos los casos. MobileNetV2 confirma ser más ligero con un consumo de 0.90 GB, seguido de ResNet50 con un consumo de 1.13 GB, sin embargo, dado que todos utilizan una porción mínima de los 40 GB disponibles en la GPU la memoria no representa ser un factor diferenciador critico en el entorno de trabajo.

Mientras tanto el uso del CPU muestra en el caso de ResNet50 una eficiencia excepcional al solo utilizar un 12.9% de la capacidad del procesador, a diferencia de DenseNet73.6% y MobileNetV2 con 70.2% de consumo de CPU.

La razón del alto consumo es por las operaciones de gestión de datos más intensivas en estos modelos por parte del CPU antes de procesar la información en la GPU, demostrando que ResNet50 es un modelo de alta precisión y además una opción escalable para entornos clínicos donde los recursos de CPU pueden ser un factor limitante.

Esta información es fundamental para tomar decisiones no solo basadas en la precisión del modelo, sino también en la eficiencia computacional, especialmente en contextos donde los recursos de hardware son limitados o se requiere escalabilidad.

3.3.4. Procesamiento realizado

Para realizar el entrenamiento del modelo se realizó un preprocesamiento a las imágenes, donde se estandarizó el formato con la finalidad de mejorar la calidad del conjunto de datos y optimizar el rendimiento del modelo. El procesamiento que se aplicó consta de distintas etapas:

1. **Filtrado Monoclase y Redimensión de Imágenes**
    - **Filtrado de imágenes del conjunto seleccionado:** La selección de imágenes a utilizar se realizó mediante un filtrado con la condición de tomar imágenes que tengan solo una de las 5 clases, sin la presencia de otras patologías del dataset.
    - **Redimensión a 224x224 píxeles:** Esta resolución fue seleccionada por ser la dimensión de entrada de las arquitecturas preentrenadas que se emplearon en el presente trabajo (ResNet50, DenseNet121 y MobileNetV2), estas redes han sido entrenadas previamente sobre el conjunto de datos ImageNet con esta resolución, por tal razón mantener el mismo formato y resolución de las imágenes de entrada permite aprovechar el aprendizaje sin modificar la arquitectura de las primeras capas convolucionales. Adicional, el uso de resoluciones a mayores escalas requiere de un mayor costo computacional de procesamiento y memoria, mientras que a menor resolución ocasiona perdida de detalles que pueden ser relevantes para el trabajo como opacidades pulmonares.

2. **Balanceo de Clases mediante Aumentación de Datos (Offline)**

   El conjunto de datos filtrados presenta un notorio desbalance de clases. Por lo que se implementó distintas estrategias de aumento de datos offline con aproximadamente 10,000 imágenes por clase:

   - **Submuestreo (Subsampling):** La clase Normal se redujo de más de 60,000 a 10,000 muestras mediante una selección aleatoria.
   - **Aumento de Datos:** Las clases minoritarias (Cardiomegaly, Nodule, Pleural_Thickening, Pneumothorax) fueron aumentadas utilizando ImageDataGenerator. Se aplica transformaciones aleatorias y clínicamente coherentes para generar nuevas muestras sintéticas: rotación (±10◦), zoom (±10%), y desplazamientos horizontales y verticales (±5%) este proceso se realizó de forma offline, guardando las imágenes generadas en disco para crear un conjunto de datos base, permitiendo simular variaciones reales en las imágenes sin alterar la información clínica relevante.

3. **Transformaciones para el entrenamiento del Modelo**

   Posterior al balanceo del dataset, se realizó una seria de transformaciones aplicadas en las imágenes durante la fase de entrenamiento para mejorar la robustez y el rendimiento del modelo.

   - **Aumento de Datos en Tiempo Real (Online):** Para mejorar la generalización de los modelos y prevenir el sobreajuste se utilizó una segunda capa de aumento de datos en tiempo real, estas transformaciones se aplicaron de forma aleatoria a las imágenes del conjunto de entrenamiento en cada época, justo antes de ser utilizadas por el modelo.
     - **Transformaciones Afines:** Se realizaron rotaciones, traslaciones y escalado adicional mediante RandomAffine, técnica de aumentación geométrica encargada de aplicar transformaciones útiles para analizar imágenes médicas que rara vez están perfectamente estandarizadas.
     - **Ajuste de Color:** Variaciones aleatorias en el brillo y contraste de la imagen ControlJitter, técnica de aumentación fotométrica, encargada de modificar aleatoriamente propiedades de color y luz en la imagen, permitiendo que los modelos sean más robustos ante las diferencias en la exposición y el

procesamiento de radiografías realizadas de distintas máquinas de rayos X y centros médicos.

- **Preparación del Tensor para el Modelo:** Este proceso es un paso fundamental para la preparación de las imágenes de radiografías, requerido previo al procesamiento de las redes neuronales preentrenadas.
  - **Conversión a Tensor:** Las imágenes originales en formato PNG fueron convertidas en formato unificado compatible, estas imágenes se convierten en tensores que son matrices numéricas multidimensionales usadas por la librería PyTorch para la realización de cálculos.
  - **Normalización para Transfer Learning:** Esta etapa ajusta los valores numéricos de los pixeles de la imagen para optimizar el aprendizaje del modelo, esta fase se compone de dos etapas importantes para el correcto aprendizaje por transferencia.
    - **Escalado:** Los valores de pixeles fueron escaladas al rango de [0, 1] dividido por 255, permitiendo estandarizar el rango de entrada y estabiliza el entrenamiento del modelo.
    - **Estandarización:** Los valores se normalizan utilizando la media y la desviación estándar especificas del dataset ImageNet. Los modelos seleccionados fueron preentrenados con millones de imágenes de ImageNet, este proceso asegura que las nuevas imágenes de radiográficas tengan una distribución estadística similar a los datos con los que los modelos fueron entrenados originalmente, favoreciendo a la transferencia de conocimiento.

4. **Organización de Conjuntos de Datos:**

   Se divide los datos en tres subconjuntos de imágenes garantizando que las imágenes estén preparadas para ser interpretadas correctamente por las distintas arquitecturas.

   - 80% para el entrenamiento.
   - 10% para la validación.
   - 10% para pruebas.

3.3.5. Modelado

El trabajo busca comparar y evaluar las distintas capacidades de las arquitecturas de CNN en la tarea de clasificación multiclase de enfermedades pulmonares. Todas las arquitecturas implementadas en PyTorch y utilizando aprendizaje por transferencia.

- **Arquitecturas Seleccionadas:** ResNet50, DenseNet121 y MobileNetV2 por su popularidad en estudios previos, su disponibilidad al ser modelos preentrenados, su diversidad en profundidad, complejidad y su rendimiento computacional.
- **Aprendizaje por Transferencia y Fine-Tuning:** Todos los modelos se inicializaron con pesos preentrenados en ImageNet, se usó una estrategia de fine-tuning profundo, en el cual se congeló las capas iniciales para preservar el aprendizaje de características de bajo nivel, las capas profundas junto con el clasificador final fueron reentrenado en las 3 arquitecturas.
  - **ResNet50:** Se reentrenaron los bloques layer3 y layer4.
  - **DenseNet121:** Se reentrenaron los bloques denseblock3, y denseblock4 y la capa de normalización final.
  - **MobileNetV2:** Se reentrenaron las últimas capas desde features [14] en adelante.
- **Hiperparámetros de Entrenamiento:** Todos los modelos se entrenarán bajo los mismos hiperparámetros para evitar sesgos o comparaciones erróneas, garantizando que las diferencias observadas en los diferentes modelos se deban a su arquitectura y no a variaciones en los datos o condiciones de entrenamiento:
  - Optimizador: AdamW.
  - Función de Pérdida: CrossEntropyLoss.
  - Planificador de Tasa de Aprendizaje: CosineAnnealingLR.
  - Tamaño de Lote (Batch Size): 128.
  - Número de Épocas: 20.
  - Métricas de Evaluación: Precisión, sensibilidad, F1-Score

3.3.6. Validación

Ground Truth para Validación

En este estudio en particular, el término “ground truth” (verdad de referencia) se refiere a las etiquetas médicas asignadas a una imagen dentro del conjunto de datos utilizados, estas etiquetas son consideradas como correctas y se utilizan para calcular la efectividad del modelo, comparándolas con las predicciones realizadas.

- **En el conjunto ChestX-ray14:** Para este conjunto de datos, las etiquetas ya habían sido generadas previamente, aplicado sobre los informes médicos redactados por radiólogos. A pesar de que este método puede contener cierto margen de error, ha sido ampliamente utilizado en la literatura científica y se ha utilizado como base para entrenar números modelos reconocidos. Cada imagen puede estar asociada a una o más enfermedades (lo que se conoce como clasificación multietiqueta), sin embargo, en este trabajo se adaptó el problema a una clasificación multiclase, seleccionando imágenes con una sola etiqueta.
- **Validación supervisada:** Durante el entrenamiento y evaluación del modelo, se utilizarán las etiquetas diagnósticas como valores de referencia para comparar con las predicciones generadas por el modelo. En cada iteración, se verifica si la clase predicha coincide con la clase real asignada en el conjunto de datos, lo que permite calcular métricas de rendimiento como precisión y sensibilidad. De este modo, se puede contrastar el rendimiento del modelo sobre un conjunto clínicamente validado, fortaleciendo la confiabilidad del proceso de validación.

Métricas para Evaluación Cuantitativa

La evaluación del rendimiento de un modelo de clasificación, en el ámbito médico no debe centrarse únicamente en la precisión general. En entornos clínicos, es fundamental analizar otras métricas que reflejen la capacidad del modelo para detectar correctamente enfermedades (sensibilidad) y ofrecer un diagnóstico equilibrado, especialmente en los casos en los cuales existen clases desbalanceadas.

Para este estudio se utilizarán las siguientes métricas cuantitativas, que son ampliamente reconocidas en el ámbito del diagnóstico asistido por inteligencia artificial, las cuales fueron descritas en el capítulo 2.

- Precisión (Accuracy).
- Sensibilidad (Recall).
- F1-score.
- Matriz de Confusión.

Todas estas métricas se calcularán sobre el conjunto de prueba una vez finalizado el entrenamiento del modelo. Además, se evaluará el rendimiento por clase y de forma global, con el fin de detectar posibles sesgos del modelo hacia determinadas patologías o clases mayoritarias.

Métricas para Evaluación Cualitativa (Explicabilidad)

Aunque el presente trabajo no incluye una validación directa por parte de especialistas médicos, se incorpora una evaluación subjetiva visual mediante técnicas de explicabilidad. Esto con el fin de verificar la coherencia clínica de las predicciones del modelo.

- **Uso de Grad-CAM:** Se aplicará la técnica Grad-CAM (Gradient-weighted Class Activation Mapping) para generar mapas de activación en las imágenes clasificadas correctamente y algunas clasificadas incorrectamente.
  Esta técnica permite visualizar qué regiones de la imagen influyeron más en la decisión del modelo, proporcionando así una explicación visual de sus predicciones.
- **Criterio de coherencia clínica:** Evaluaremos cualitativamente los mapas generados en función de:
  - Si las regiones activadas por el modelo coinciden con zonas anatómicas relevantes para la enfermedad.
  - Si en los casos positivos (por ejemplo, Pneumotorax o Nódulo), las zonas activadas corresponden a opacidades visibles, zonas densas o anomalías esperadas en una lectura médica convencional.
  - Si en las imágenes etiquetadas como “normal”, las regiones activadas no presentan patrones erráticos o activaciones sin justificación clínica.

Aunque esta validación no reemplaza el juicio de un profesional médico, permite aportar una capa adicional de confianza sobre el comportamiento interno del modelo, favoreciendo su transparencia y aproximación clínica. Este tipo de evaluación es ampliamente utilizado en estudios donde se aplican modelos de visión por computador a tareas médicas, como apoyo a la explicabilidad del sistema.

3.3.7. Despliegue

Aunque no se realiza una implementación en producción real:

- Se discute su aplicación en entornos clínicos reales.
- Se plantea que el modelo sea replicable, accesible, y eficiente para hospitales con recursos limitados.
- Grad-CAM sirve como mecanismo de confianza para su futura adopción en la práctica médica.

4. Planteamiento de la comparativa

En el presente capítulo se describe el planteamiento del trabajo desarrollado para evaluar el rendimiento de diferentes modelos de aprendizaje profundo aplicado a la tarea de clasificación multiclase de enfermedades pulmonares de imágenes de radiografías de tórax. El objetivo de esta comparativa es determinar cuál de las arquitecturas seleccionadas (ResNet50, DenseNet121 y MobileNetV2) ofrece un mejor equilibrio entre precisión en el diagnóstico, eficiencia computacional y su capacidad de explicabilidad visual mediante técnicas como Grad-CAM.

Estas arquitecturas, han sido utilizadas en tareas de clasificación de imágenes en el ámbito médico, las cuales serán entrenadas bajo iguales condiciones y conjunto de datos públicos, Esto permite realizar una comparación imparcial y replicable para cada arquitectura. El estudio se centra en la identificación automatizada de cinco condiciones clínicas frecuentes: cardiomegalia, engrosamiento pleural, neumotórax, nódulo pulmonar y normalidad, a partir de radiografías de tórax.

El capítulo describe el contexto del problema clínico y la importancia de contar con herramientas de apoyo para el diagnóstico por imágenes, se detallan las soluciones a evaluar, los criterios definidos para la comparación. Finalmente, se detalla y describe los resultados con el fin de garantizar su validez.

4.1. Problema clínico y contexto de aplicación

En el presente capítulo se aborda el problema de clasificación multiclase de enfermedades pulmonares en radiografías de tórax, centrándose en aquellas patologías cuya diferenciación por imagen resulta especialmente compleja debido a similitudes visuales. La naturaleza sutil de los patrones radiológicos de enfermedades como cardiomegalia, engrosamiento pleural, neumotórax y nódulo pulmonar impone retos técnicos al momento de evaluar el rendimiento de modelos de aprendizaje profundo.

La comparativa propuesta considera que, aunque existen múltiples modelos capaces de abordar tareas de clasificación, sus capacidades pueden variar significativamente en función de la arquitectura utilizada, el tipo de preentrenamiento aplicado, la sensibilidad ante clases

desbalanceadas y la interpretabilidad de los resultados. Por tanto, se plantea un análisis estructurado que no solo evalúa el desempeño predictivo, sino también la robustez y utilidad práctica de cada modelo.

4.2. Soluciones para comparar

El presente estudio compara tres arquitecturas de redes neuronales convolucionales reconocidas por su eficacia en tareas de clasificación de imágenes como se detalló en el capítulo 2.5, las arquitecturas DenseNet121, ResNet50 y MobileNetV2 han sido seleccionadas debido a su solidez técnica, disponibilidad como modelos preentrenados y su gran desempeño en aplicaciones médicas como el análisis de imágenes radiológicas.

Las tres arquitecturas emplearan el esquema de transfer learning, utilizando pesos preentrenados en el conjunto de datos ImageNet, esta estrategia permite aprovechas la información aprendida por las redes y emplearlas en la tarea de clasificación multiclase de enfermedades pulmonares.

- ResNet50
- DenseNet121
- MobileNetV2

Estas arquitecturas buscan implementar un equilibro entre precisión, complejidad y eficiencia computacional, lo que las convierte en arquitecturas útiles para evaluar el desempeño en el ámbito médico. La comparación se realizará bajo condiciones controladas, asegurando que cada modelo sea evaluado con el mismo conjunto de datos, parámetros de entrenamiento y métricas de análisis.

4.3. Criterios de Comparación

Para evaluar de forma objetiva el rendimiento de los modelos entrenados con cada arquitectura se ha definido una serie de criterios cuantitativos y cualitativos que permitirán comparar su capacidad de diagnosticar cada patología, su eficiencia computacional y su nivel de explicabilidad. Estos criterios han sido seleccionados en base a criterios y estándares del ámbito médico y los objetivos específicos del presente trabajo de fin de máster. Los criterios son:

4.3.1. Métricas cuantitativas del rendimiento

Se calculan a partir de las predicciones del modelo y las etiquetas reales del conjunto de datos de prueba:

- Precisión (Accuracy).
- Sensibilidad (Recall).
- F1-score.

4.3.2. Eficiencia Computacional

Se evaluará la viabilidad técnica de los modelos permitiendo valorar el costo computacional asociado al uso real de cada modelo, teniendo relevancia en aplicaciones con recursos limitados mediante:

- Tiempo de entrenamiento por época
- Tiempo promedio de inferencia por imagen
- Uso de memoria (RAM y VRAM) durante el entrenamiento

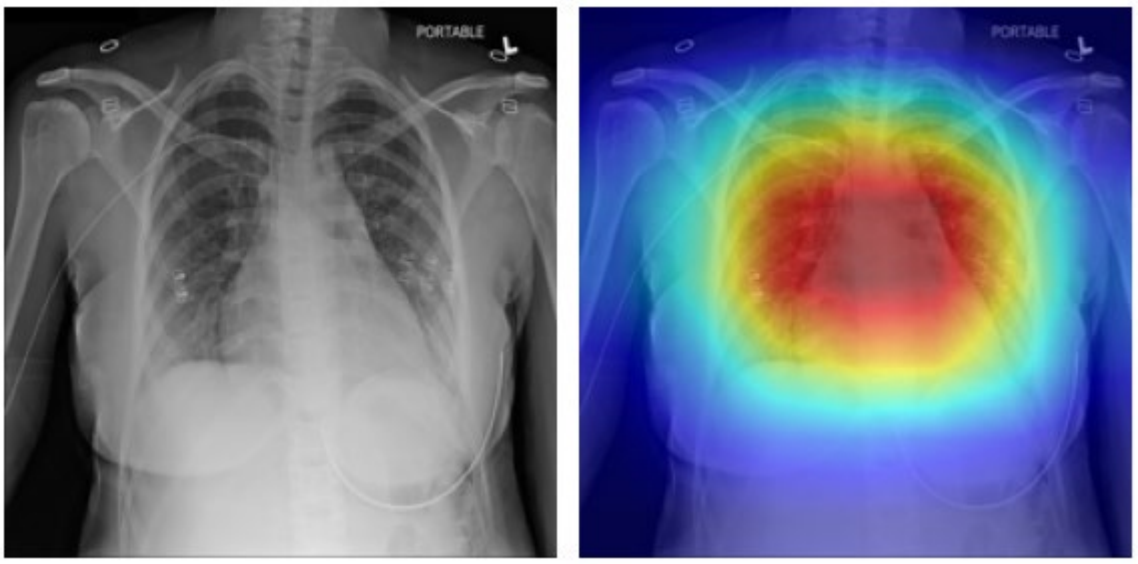
4.3.3. Explicabilidad visual

Se evaluará la capacidad del modelo para justificar sus decisiones mediante visualizaciones generadas con Grad-CAM, donde, se analizará si las regiones activadas por cada modelo corresponden con zonas anatómicas importantes en las radiografías de tórax. Se incluirán casos clasificados positivos y negativos para analizar coherencia de las activaciones.

Este conjunto garantiza una evaluación integral permitiendo asegurar la aplicabilidad en trabajos reales y su replicabilidad en un entorno médico.

La **Figura 14** muestra un ejemplo de uso de Grad-CAM aplicada a una radiografía clasificada como Cardiomegalia donde se observa las zonas de activación para su uso clínico.

Figura 14. *Resultado de Aplicar Grad-CAM a imagen con cardiomegalia.*



Fuente: Elaboración propia

Este trabajo busca garantizar una comparativa sólida y equitativa entre las soluciones propuestas, controlando las variables técnicas y permitiendo que las diferencias observadas sean atribuibles exclusivamente al desempeño de cada arquitectura.

5. Desarrollo de la comparativa

En el siguiente capítulo se presenta los resultados numéricos obtenidos por cada arquitectura en el conjunto de prueba. En cada modelo analiza el reporte de clasificación, la matriz de confusión y las curvas de aprendizaje para obtener una visión completa del rendimiento en el diagnóstico y su comportamiento durante el entrenamiento.

5.1. Análisis de Clasificación

En este apartado se presentan y comparan los resultados detallados de las tres arquitecturas en el conjunto de prueba. El análisis se centra en las métricas de precisión, sensibilidad (recall), F1-score y para obtener una visión completa del rendimiento diagnóstico de cada modelo y su comportamiento específico ante cada patología.

Los resultados comparativos se presentan en la **Tabla 4,** donde se muestran los resultados de cada arquitectura y su precisión global correspondiente.

Tabla 4. *Tabla Comparativa de Métricas de Rendimiento por Arquitectura*

		Normal	Cardiomegalia	Neumotórax	Nódulo Pulmonar	Engrosamiento Pleural
ResNet50	**precision**	0.8176	0.9743	0.9608	0.9189	0.9771
Accuracy Global: 92.92%	**recall**	0.8470	0.9860	0.9570	0.8730	0.9830
	f1-score	0.8320	0.9801	0.9589	0.8954	0.9801
DenseNet121	**precision**	0.8239	0.9612	0.9515	0.9348	0.9705
Accuracy Global: 92.82%	**recall**	0.8420	0.9900	0.9610	0.8600	0.9880
	f1-score	0.8328	0.9754	0.9562	0.8958	0.9792

MobileNetV2	**precision**	0.6971	0.8725	0.8207	0.7181	0.7571
Accuracy Global: 77.24%	**recall**	0.8470	0.9310	0.7280	0.5580	0.7980
	f1-score	0.7648	0.9008	0.7716	0.6280	0.7770

Fuente: Elaboración propia

En la **Tabla 4** se observa que los modelos ResNet50 y DenseNet121 presentan una clara superioridad, donde ResNet50 obtuvo una precisión global de 92.92% seguido de DenseNet121 con una precisión global de 92.82% demostrando una alta eficacia en la definición e identificación de enfermedades. El modelo MobileNetV2 obtuvo una precisión global inferior de 77.24%.

Las características destacables de las 3 arquitecturas muestran patrones visuales definidos para las enfermedades Cardiomegalia y Engrosamiento Pleural, donde destaca ReseNet50 con valores de F1-Score superiores para Cardiomegalia 98.01% y Engrosamiento Pleural 98.01% evidenciando una robustez sobresaliente en estas clases.

Las clases que presentaron mayor dificultad para identificar fueron Normal y Nódulo Pulmonar, siendo en este caso superior el modelo DenseNet121 con mejores resultados en F1-Score tanto en Normal con 83.28% y Nódulo Pulmonar con 89.58% sugiriendo que esta arquitectura puede ser mejor para diferenciar casos sanos y detectar patrones sutiles.

El análisis del Modelo MobileNetV2 muestra un rendimiento bajo, teniendo resultados problemáticos como el caso de Nódulo Pulmonar con un F1-Score de 62.80% y un Recall de 55.80% demostrando que el modelo es incapaz de detectar la mitad de los nódulos reales, siendo un caso crítico y de cuidado en su uso para el contexto clínico.

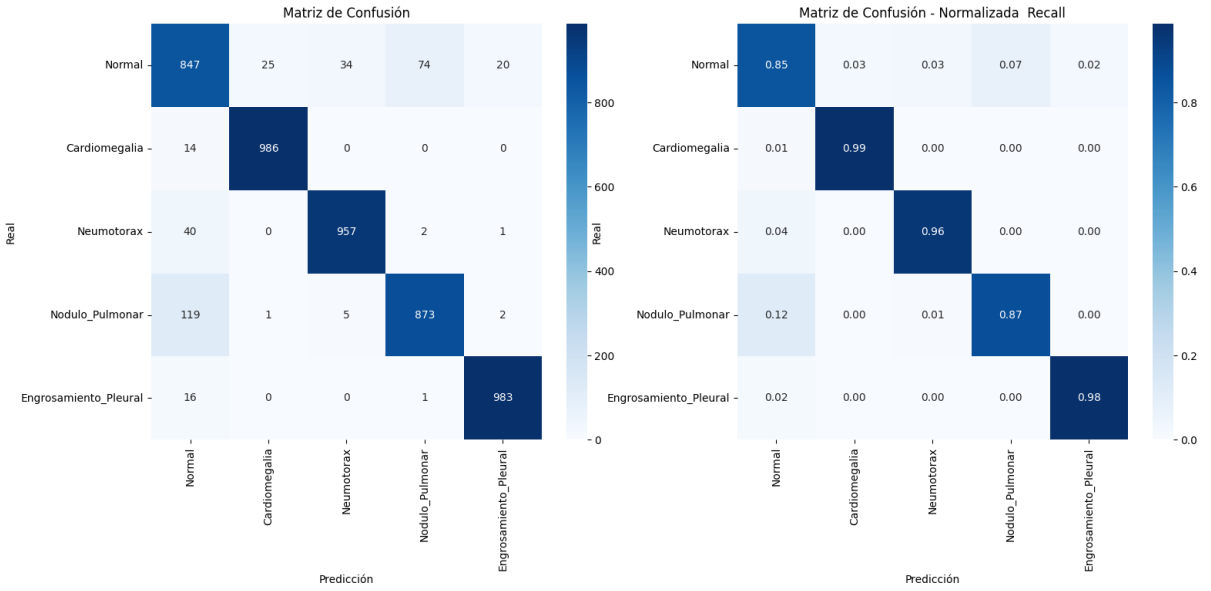
El análisis demuestra que ResNet50 obtiene la mayor precisión global por un mínimo margen de diferencia, DenseNet121 presenta resultados más confiables siendo un modelo más equilibrado especialmente en clases que requieren una mayor detección de patrones sutiles, demostrando que ambos Modelos son una solución de alto rendimiento para la clasificación y detección de enfermedades. El modelo MobileNetV2 queda descartado como una opción

clínicamente viable debido a su bajo rendimiento general y fallo critico en la detección de nódulos.

5.2. Análisis Detallado de las Matrices de Confusión

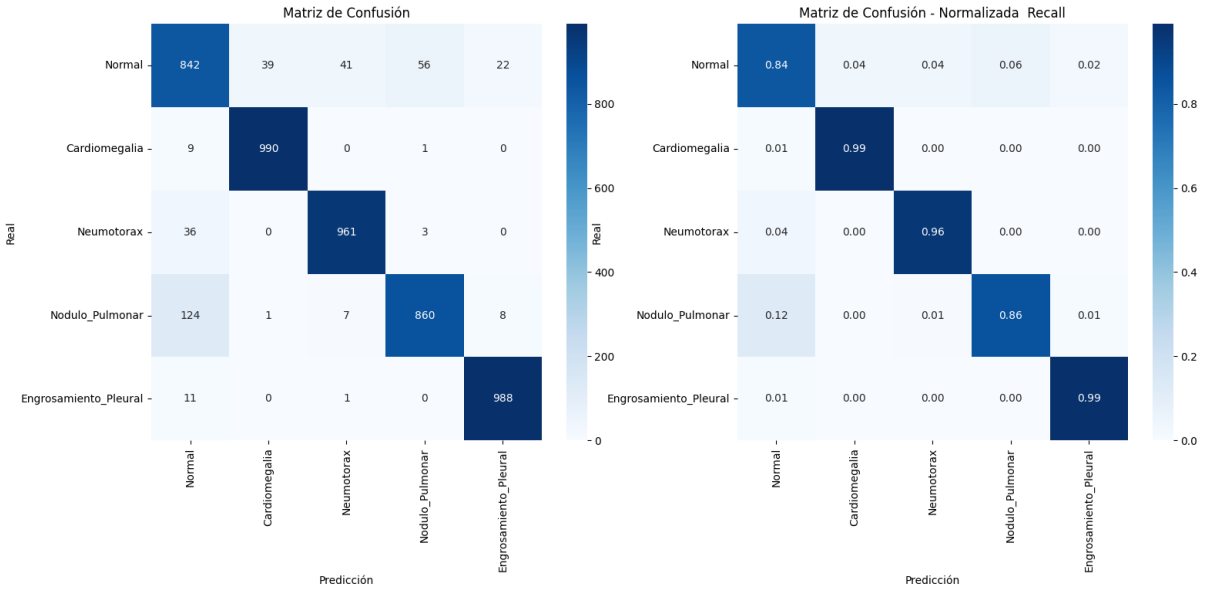
El análisis de las matrices de confusión permite conocer visualmente los patrones de acierto y error de cada arquitectura, logrando evidenciar como los modelos procesan la información sus errores y aciertos en cada clase.

Figura 15. *Matriz de Confusión ResNet50.*



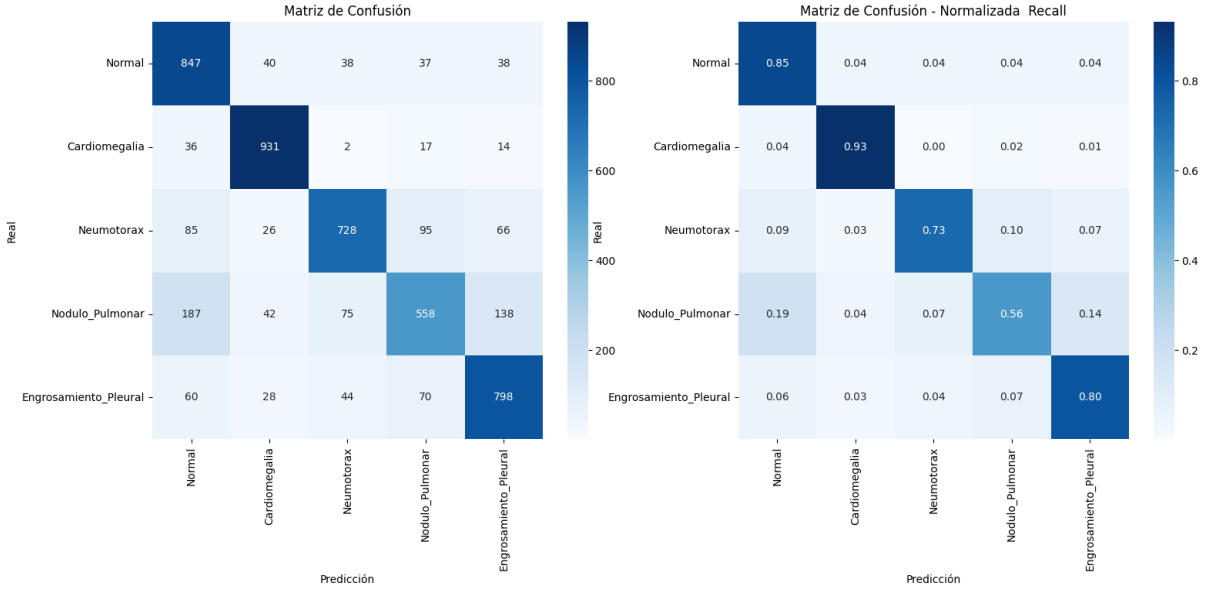
Fuente: Elaboración propia

Figura 16. *Matriz de Confusión DenseNet121*



Fuente: Elaboración propia

Figura 17. *Matriz de Confusión MobileNetV2*



Fuente: Elaboración propia

Los modelos ResNet50 y DenseNet121 demuestran ser eficaces en la detección y la clasificación de las clases. Como se muestra en la matriz de confusión de **Figura 15** y en **Figura 16** ambos modelos clasificaron la mayoría de los casos, destacando en las clases de Cardiomegalia con 986 aciertos ResNet50 y 990 aciertos DenseNet121, Engrosamiento Pleural

983 aciertos Resnet50 y 988 aciertos DenseNet, demostrando excelentes resultados para aprender patrones a gran escala.

El modelo MobileNetV2 a pesar de su bajo desempeño y rendimiento en las distintas clases **Figura 17** muestra buenos resultados en Cardiomegalia con 931 aciertos, aunque su rendimiento decae en las distintas patologías.

Se puede observar que los modelos disponen de ciertas limitaciones, siendo una tarea de mayor dificultad para los tres modelos la diferenciación de la clase Normal y Nódulo, siendo un resultado a tomar en cuenta, en el caso de ResNet50 detecta 74 imágenes Normales como Nódulo Pulmonar mientras que DenseNet121 clasificó 124 imágenes como Normal, esto indica que DenseNet121 es más ligeramente más propenso a pasar por alto un Nódulo Pulmonar en comparación con ResNet50.

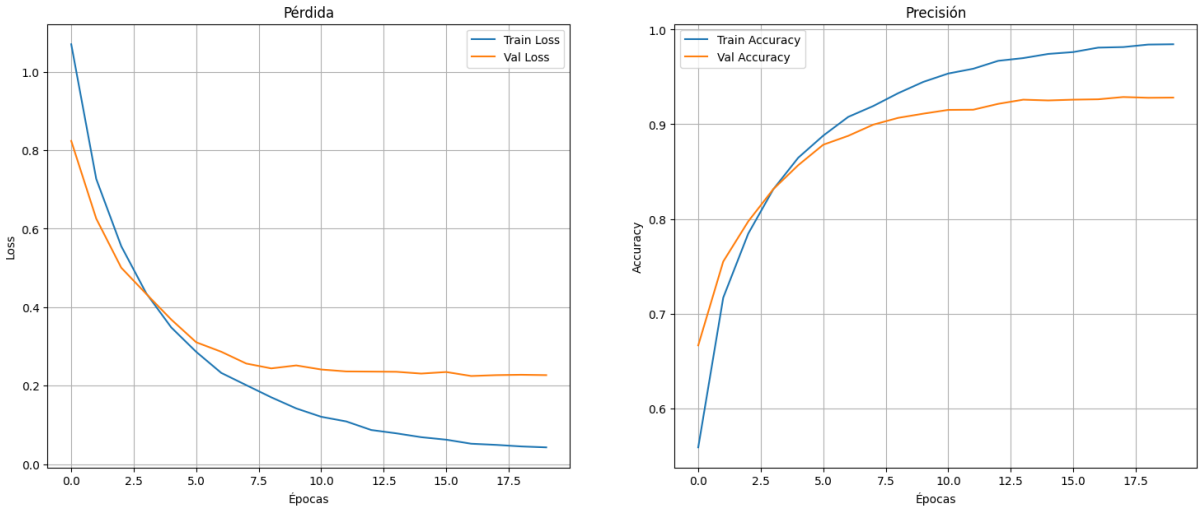
En el caso de MobileNetV2 se evidencia errores graves al omitir 187 casos de Nódulo Pulmonar como Normal y 138 Nódulos Pulmonares como Engrosamiento Pleural demostrando una incapacidad de aprender características y patrones discriminativas robustas en estas enfermedades.

Las matrices de confusión aportan un mejor entendimiento de los resultados, demostrando que ResNet50 y DenseNet121 son modelos robustos los cuales presentan complicaciones al diferenciar ciertos casos como nódulos pulmonares, en este caso DenseNet121 es presenta un mejor equilibro al no generar falsos positivos de Nódulos pulmonares, mientras que ResNet50 destaca al no omitirlos. La matriz de MobileNetV2 confirma que no es una arquitectura fiable para esta tarea por su alta tase de falsos positivos y datos no generalizados.

5.3. Análisis de las Curvas de Aprendizaje

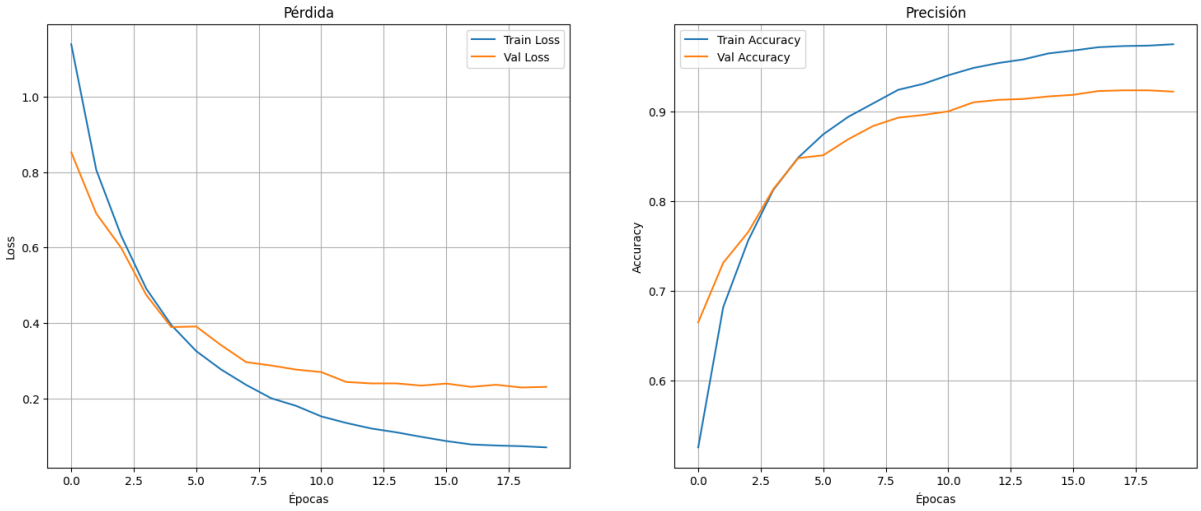
El análisis de las curvas de pérdida (loss) y precisión (accuracy) durante las 20 épocas de entrenamiento permite evaluar la calidad del proceso de aprendizaje de cada modelo, valorar la estabilidad de la convergencia y la capacidad de generalización del modelo, identificando posibles signos de sobreajuste.

Figura 18. *Curva de Entrenamiento ResNet50*



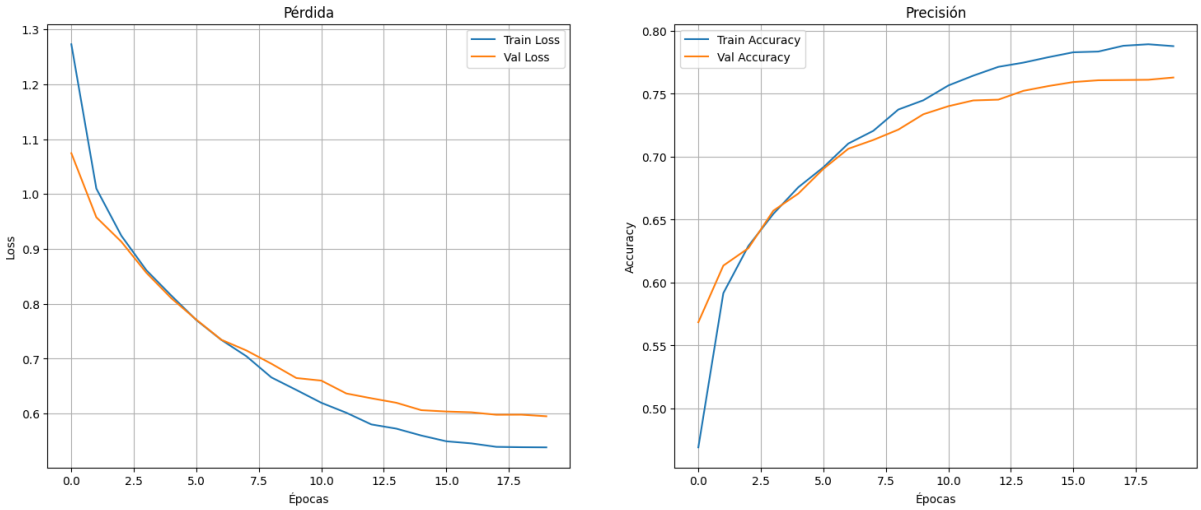
Fuente: Elaboración propia

Figura 19. *Curva de Entrenamiento DenseNet121*



Fuente: Elaboración propia

Figura 20. *Curva de Entrenamiento MobileNetV2*



Fuente: Elaboración propia

Las curvas del modelo ResNet50 en la **Figura 18** muestran un proceso de entrenamiento similar al modelo DenseNet121 en la **Figura 19**. Muestran una correcta estabilización progresiva, demostrando que los modelos aprendieron patrones del conjunto de datos de forma eficiente. Ambos modelos alcanzaron su mejor rendimiento en las etapas finales del entrenamiento, ResNet50 alcanzó una precisión de 92.64% en la capa 17, mientras que DenseNet121 alcanzó la mejor precisión en la época 19 con un 92.32%, demostrando que el ciclo de 20 épocas fue adecuado para que ambos modelos alcanzaran su máximo potencial.

Ambos modelos muestran una excelente generalización, donde ResNet50 muestra una pérdida de validación de 22.48% mientras que DenseNet121 una pérdida de validación de 22.90% demostrando bajos valores que demuestra que el modelo no ha sufrido sobreajuste y a generalizado adecuadamente cada época.

El modelo MobileNetV2 muestra en la **Figura 20** un claro sobreajuste, a partir de la época 10 se observa una divergencia clara y creciente entre las curvas. Se puede evidenciar que la perdida de validación se estanca y tiende a aplanarse, indicando sobreajuste, el modelo comienza a memorizar patrones del conjunto de entrenamiento como consecuencia su bajo rendimiento en el conjunto de pruebas.

El análisis de las curvas de aprendizaje permite contrastar los resultados cuantitativos, demuestra que ResNet50 y DenseNet121 son más precisas y disponen de procesos de entrenamiento más robustos y una mejor generalización, a diferencia del bajo rendimiento de MobileNetV2 se debe a su sobreajuste, demostrando ser la arquitectura menos adecuada.

5.4. Evaluación con GRAD-CAM

En el presente punto se implementó la técnica de explicabilidad GRAD-CAM, el objetivo de este análisis cualitativo es verificar que en los aciertos los modelos se enfoquen en regiones clínicamente relevantes, adicional, diagnosticar patrones de errores al analizar por qué se producen las clasificaciones incorrectas.

Los resultados de esta evaluación se presentan en la **Tabla 5,** en la cual se presenta la presenta la tasa de predicción de acierto por cada clase según cada arquitectura. El objetivo de esta tabla es visualizar si los modelos activan correctamente zonas anatómicas en función del diagnóstico real. Las imágenes generadas por Grad-CAM se adjuntan en los **Anexo C.**

Tabla 5. *Análisis Cualitativo con Grad-CAM*

		ResNet50	DenseNet121	MobileNetV2
Normal	**Normal**	100.00%	100.00%	70.80%
	Cardiomegalia	0.00%	0.00%	15.90%
	Neumotórax	0.00%	0.00%	0.70%
	Nódulo Pulmonar	0.00%	0.00%	11.70%
	Engrosamiento Pleural	0.00%	0.00%	0.90%

Cardiomegalia	**Normal**	0.00%	0.00%	0.00%
	Cardiomegalia	100.00%	100.00%	98.70%
	Neumotórax	0.00%	0.00%	0.50%
	Nódulo Pulmonar	0.00%	0.00%	0.40%
	Engrosamiento Pleural	0.00%	0.00%	0.30%
Neumotórax	**Normal**	0.00%	0.00%	0.00%
	Cardiomegalia	0.00%	0.00%	0.00%
	Neumotórax	100.00%	100.00%	98.70%
	Nódulo Pulmonar	0.00%	0.00%	1.20%
	Engrosamiento Pleural	0.00%	0.00%	0.10%
Nódulo Pulmonar	**Normal**	0.00%	0.00%	0.40%
	Cardiomegalia	0.00%	0.00%	0.10%
	Neumotórax	0.00%	0.00%	0.80%
	Nódulo Pulmonar	100.00%	100.00%	97.60%
	Engrosamiento Pleural	0.00%	0.00%	1.10%

Engrosamiento Pleural	**Normal**	0.00%	0.00%	0.10%
	Cardiomegalia	0.00%	0.00%	5.30%
	Neumotórax	0.00%	0.00%	1.30%
	Nódulo Pulmonar	0.00%	0.00%	33.30%
	Engrosamiento Pleural	100.00%	100.00%	60.10%

Fuente: Elaboración propia

Las predicciones correctas de los modelos mejor puntuados como **DenseNet121** y **ResNet50** demostraron resultados sobresalientes en la identificación y localización de patrones visuales característicos de cada enfermedad.

Los resultados delimitan correctamente los mapas de activación de forma precisa en base a la información clínica que proporciona las imágenes, se muestra una fuerte correlación entre la calidad de los mapas de activación y el rendimiento cuantitativo de la tabla.

El modelo que usa la arquitectura MobileNetV2 obtuvo los resultados más bajos, generó mapas de calor débiles y menos focalizados**.** Esto evidencia que su limitada capacidad afecta directamente en la detección y localización de zonas anatómicas de importancia por sus bajas métricas y menor precisión.

La evaluación mediante Grad-CAM permite validar el éxito para los modelos con ResNet50 y DenseNet121 los cuales muestran un aprendizaje de características visuales clínicamente relevantes, sustentando la confianza en sus resultados, obteniendo un diagnóstico preciso de sus limitaciones. Además, la tabla permite conocer las deficiencias presentadas en el Modelo MobileNetV2 concluyendo que las arquitecturas profundas fueron superiores en las métricas y la coherencia visual de los resultados proporcionados por Grad-CAM.

5.4.1. Análisis de Errores con Grad-CAM.

El análisis de los mapas de Activación en casos de error permite conocer las limitaciones prácticas de cada modelo, este análisis selecciona dos casos de prueba donde el modelo

MobileNetV2 cometió errores importantes y se evalúa el comportamiento de las tres arquitecturas. Los resultados de esta comparación se presentan en la **Tabla 6**.

Tabla 6. *Análisis comparativo de Errores con Grad-CAM*

		ResNet50	DenseNet121	MobileNetV2
Normal	**Normal**	12.90%	0.00%	48.40%
	Cardiomegalia	0.00%	0.00%	0.50%
	Neumotórax	0.00%	0.00%	19.40%
	Nódulo Pulmonar	87.10%	100.00%	30.20%
	Engrosamiento Pleural	0.00%	0.00%	1.50%
Neumotórax	**Normal**	3.20%	100.00%	56.80%
	Cardiomegalia	0.10%	0.00%	0.40%
	Neumotórax	96.70%	0.00%	22.00%
	Nódulo Pulmonar	0.00%	0.00%	20.10%
	Engrosamiento Pleural	0.00%	0.00%	0.70%

Fuente: Elaboración propia

El primer caso corresponde a una imagen cuya etiqueta real es Normal, pero fue clasificada erróneamente, en el modelo ResNet50 clasificó como Nódulo Pulmonar con una confianza del 87.10% y DenseNet121 clasificó como Nódulo Pulmonar con una confianza del 100%. Ambos casos sugieren un fallo en las dos arquitecturas, demostrando que son propensas a malinterpretar estructuras anatómicas ambiguas. En el caso del Modelo MobileNetV2 clasificó

correctamente Normal con una confianza de 48.40% la cual refleja su limitada capacidad discriminativa.

En la segunda prueba se evaluó una imagen con etiqueta real Neumotórax, en el caso de ResNet50 realizó una predicción correcta identificando la clase con una confianza del 96.70% demostrando la robustez para detectar esta patología, a pesar de ser un caso complicado de identificar en otras arquitecturas.

A pesar de que el mejor modelo en base a sus resultados sea DenseNet121 fallo en esta prueba al clasificar la imagen como normal con un 100% de confianza, mostrando casos que el modelo falla en la detección y diagnóstico oportuno de las patologías. Mientras que MobileNetV2 fallo al clasificar como normal la imagen con una confianza de 56.80%.

Este apartado permite conocer que, aunque los modelos ResNet50 y DenseNet121 son altamente precisos, pueden cometer errores con una confianza muy alta. Estos resultados son importantes para entender los límites y el uso seguro de estas tecnologías como herramientas de apoyo al diagnóstico. Estas imágenes se adjuntan en el **Anexo D.**

6. Discusión y análisis de resultados

La discusión de resultados constituye una etapa crítica en la que se contextualizan, interpretan y contrastan los hallazgos del presente trabajo con estudios previos. En este caso, se ha realizado una clasificación multiclase de enfermedades torácicas usando redes neuronales convolucionales profundas. Los modelos entrenados —DenseNet121, MobileNetV2 y ResNet50— presentaron un rendimiento variable, siendo ResNet50 el más robusto en métricas como precisión y F1-score junto con DenseNet121. Esto sugiere que las conexiones densas permiten una mejor propagación de características entre capas, favoreciendo el aprendizaje de patrones visuales complejos.

Más allá de las métricas globales, se analizaron los valores por clase, revelando que ciertas patologías como normalidad o nódulo resultaron más difíciles de clasificar correctamente. Estos hallazgos no solo reflejan el comportamiento del modelo, sino también la calidad de las etiquetas y la complejidad inherente de las imágenes radiológicas. En este contexto, los errores frecuentes observados en las matrices de confusión fueron fundamentales para guiar el análisis cualitativo.

Asimismo, la generación de mapas Grad-CAM permitió verificar visualmente si las decisiones del modelo eran coherentes con los patrones clínicos. Esta combinación de análisis cuantitativo y visual sustenta una evaluación integral de los resultados y respalda su interpretación dentro del contexto médico real. Los valores generales de desempeño de cada clase y modelo se presentan en la **Tabla 7**.

Tabla 7. *Resultados Generales.*

Clase	Métrica	DenseNet121	ResNet50	MobileNetV2
Normal	Precision	0.8239	0.8176	0.6971
	Recall	0.8420	0.8470	0.8470
	F1-score	0.8328	0.8320	0.7648
Cardiomegalia	Precision	0.9612	0.9743	0.8725
	Recall	0.9900	0.9860	0.9310
	F1-score	0.9754	0.9801	0.9008
Neumotórax	Precision	0.9515	0.9608	0.8207
	Recall	0.9610	0.9570	0.7280
	F1-score	0.9562	0.9589	0.7716
Nódulo	Precision	0.9348	0.9189	0.7181
	Recall	0.8600	0.8730	0.5580
	F1-score	0.8958	0.8954	0.6280
Engrosamiento Pleural	Precision	0.9705	0.9771	0.7571
	Recall	0.9880	0.9830	0.7980
	F1-score	0.9792	0.9801	0.7770

Fuente: Elaboración propia

Tras la exposición de los resultados en el capítulo anterior, este apartado se dedica a su interpretación analítica. Se realiza una discusión comparativa del rendimiento de las tres arquitecturas, se profundiza en las fuentes de error comunes, se valora el papel de la explicabilidad visual y se reconocen las limitaciones inherentes al estudio, todo ello en el contexto del estado del arte.

6.1. Análisis Comparativo del Rendimiento de las Arquitecturas

El análisis comparativo entre DenseNet121, MobileNetV2 y ResNet50 permitió evaluar la capacidad diferencial de cada arquitectura para abordar la clasificación de enfermedades torácicas. ResNet50 destacó ligeramente en precisión y F1-score en comparación con DenseNet121, especialmente en clases clínicamente difíciles como 'Cardiomegalia' y 'Engrosamiento Pleural'. Su arquitectura con conexiones densas facilita la reutilización de características, lo que resulta útil en imágenes donde los patrones visuales son sutiles.

MobileNetV2, en cambio, mostró un rendimiento más limitado, aunque destaca por su eficiencia computacional. Es una arquitectura optimizada para dispositivos móviles y tareas con recursos restringidos, lo que explica su menor desempeño en clases con características menos definidas. Sin embargo, en clases con patrones más evidentes, su rendimiento fue aceptable.

DenseNet121, se ubicó entre los dos extremos, con un balance adecuado entre rendimiento y complejidad computacional. Su diseño con bloques residuales mejora el flujo de gradientes, pero podría ser comparable con ResNet50 para tareas donde se requiere la captura de detalles muy finos.

El análisis por clase mostró que los modelos tienden a tener un rendimiento más alto en clases frecuentes (como “Cardiomegalia” y “Engrosamiento Pleural”), mientras que fallan en clases menos representadas. Esto refuerza la importancia de un dataset balanceado y la necesidad de técnicas de compensación por desbalance de clases.

Los resultados globales obtenidos por cada arquitectura se presentan en la **Tabla 8,** donde se muestra las precisiones alcanzadas en el conjunto de prueba.

Tabla 8. *Precisión Global.*

Modelo	Accuracy global
DenseNet121	92.82%
ResNet50	92.92%
MobileNetV2	77.24%

Fuente: Elaboración propia

Como se observa en la **Tabla 8,** los resultados cuantitativos posicionan a:

- DenseNet121 y ResNet50 como las arquitecturas claramente superiores para esta tarea de clasificación multiclase, con precisiones globales del 92.82% y 92.92%, respectivamente. La similitud en DenseNet121 puede atribuirse a su diseño fundamental, que promueve una reutilización de características más intensiva a través de sus bloques densos. Esta conectividad directa entre capas facilita un mejor flujo de información y de gradientes, lo que resulta especialmente eficaz para aprender los patrones jerárquicos y a menudo sutiles presentes en las imágenes radiológicas.
- MobileNetV2 (77.24% de precisión) pone de manifiesto el compromiso entre eficiencia y capacidad de representación. Aunque fue la arquitectura más rápida en el entrenamiento, su diseño ligero, optimizado para entornos con recursos computacionales limitados, carece de la complejidad necesaria para capturar la alta variabilidad de las patologías estudiadas. En el contexto de una aplicación de diagnóstico clínico, donde la fiabilidad es primordial, la drástica reducción en la precisión diagnóstica hace que el ahorro computacional de MobileNetV2 no sea un intercambio aceptable.

6.2. Interpretación de los Errores de Clasificación Comunes

Los errores de clasificación observados en las matrices de confusión revelan importantes aspectos sobre el comportamiento de los modelos. Por ejemplo, se observó una alta tasa de confusión entre las clases 'Normal' y 'Nódulo Pulmonar'. Esta similitud radiológica entre ambas condiciones, sumada a la limitada resolución de las imágenes, puede explicar la dificultad del modelo para distinguirlas.

En el caso del neumotórax, una patología que puede manifestarse con signos sutiles o apenas visibles en imágenes radiológicas, se registraron varios falsos negativos. Es decir, imágenes con neumotórax fueron clasificadas como normales. Esta situación es clínicamente crítica, ya que puede llevar a una omisión diagnóstica en la práctica médica. El modelo tiende a ser conservador cuando no hay patrones claramente distinguibles.

Otro patrón relevante fue la tendencia de MobileNetV2 a etiquetar erróneamente imágenes con patologías como “Normal”. Esto puede deberse a su menor profundidad y capacidad de extracción de características, lo que lo hace más susceptible a errores en imágenes complejas. Por su parte, DenseNet121 mostró un comportamiento más consistente, con errores más reducidos y en menor cantidad.

Estos errores sugieren la necesidad de revisar tanto el modelo como la calidad de los datos y explorar posibles mejoras como la aplicación de técnicas de data augmentation más sofisticadas, entrenamiento con focal loss o incorporación de segmentación previa.

La **Tabla 9** resume los errores más frecuentes por clase, los modelos afectados y las posibles causas que explican dichos errores.

Tabla 9. *Errores comunes.*

Clase	Error más frecuente	Modelos afectados	Posible razón del error
Normal	Confundida con Nódulo Pulmonar	MobileNetV2 DenseNet121 ResNet50	Similitud visual en radiografías (zonas sin anomalías evidentes).
Cardiomegalia	Poco error, pero algunos confundidos como Normal	MobileNetV2	Dilatación cardíaca leve no detectada por redes con baja capacidad.
Neumotórax	Confundido con Normal y Nódulo Pulmonar	MobileNetV2	Neumotórax leves pueden parecer ausencia de patología.
Nódulo	Confundido con Normal y Engrosamiento Pleural	MobileNetV2	Tamaño pequeño o localización poco visible del nódulo.
Engrosamiento Pleural	Confundido con Nódulo Pulmonar y Normal	MobileNetV2	Similitud de densidades pleurales y bordes no bien diferenciados.

Fuente: Elaboración propia

El análisis detallado de las matrices de confusión de los modelos de mayor rendimiento de la revela un patrón de error consistente y clínicamente significativo:

La dificultad para distinguir entre las clases Normal y Nódulo. Esta confusión es radiológicamente coherente. Como se describe en el estado del arte, los nódulos pulmonares pueden ser de tamaño muy reducido (< 3 mm) y presentar un contraste bajo, lo que dificulta su diferenciación de estructuras vasculares o artefactos de la imagen, incluso para un radiólogo experto.

El hecho de que los modelos más potentes tropiecen en esta área sugiere que el problema roza los límites de la información discernible en radiografías 2D. Refuerza esta idea el pobre desempeño de MobileNetV2 en esta clase (F1-score de 0.6280), indicando que la detección de nódulos fue el reto más complejo de la tarea de clasificación.

6.3. Relevancia Clínica de la Explicabilidad con Grad-CAM

La explicabilidad es un aspecto esencial en aplicaciones médicas de inteligencia artificial. A través de Grad-CAM, se generaron mapas de activación que permiten visualizar en qué zonas de la imagen se enfocó el modelo al tomar una decisión. Esta funcionalidad no solo aporta transparencia, sino que también facilita una revisión crítica por parte de profesionales clínicos. En este estudio, los mapas Grad-CAM revelaron activaciones coherentes con la localización anatómica esperada. Por ejemplo, para casos de cardiomegalia, se observaron activaciones sobre el área del corazón; en neumotórax, el modelo se enfocó en los bordes pleurales. Estos resultados sugieren que el modelo aprendió a identificar patrones clínicamente relevantes, lo cual aumenta la confianza en su utilización.

Desde la perspectiva clínica, esta capacidad de explicación visual es clave para aceptar modelos de IA en hospitales. Los profesionales de salud no aceptarán una predicción sin justificación. Grad-CAM, por tanto, no solo es útil como herramienta de validación, sino también como puente entre los sistemas automatizados y el juicio clínico experto.

La explicabilidad también ayuda a detectar errores. Si un modelo toma una decisión basada en regiones irrelevantes (por ejemplo, fuera de los pulmones), esto puede alertar sobre un sobreajuste o un sesgo en el conjunto de entrenamiento.

Los beneficios observados en este estudio se presentan en la **Tabla 10,** donde se presenta las ventajas prácticas del uso de Grad-CAM con ejemplos específicos obtenidos en el desarrollo del TFM.

Tabla 10. *Relevancia Clínica.*

Beneficio clínico	Ejemplo práctico en el TFM
Validación visual por radiólogos	Verificación de que se activa la región cardiaca ante cardiomegalia
Confianza en la IA como herramienta de apoyo	Comprensión de por qué el modelo sugiere "nódulo" en un caso dado
Prevención de errores basados en sesgos	Identificación de activaciones incorrectas en regiones no patológicas
Mejora del entrenamiento médico y académico	Uso de mapas Grad-CAM como recurso didáctico

Fuente: Elaboración propia

La implementación de Grad-CAM no fue un mero ejercicio técnico, sino un paso fundamental para validar la robustez y la fiabilidad de los modelos, un objetivo clave de este TFM. Las visualizaciones generadas confirmaron de manera consistente que las predicciones de los modelos, tanto las correctas como las incorrectas, se basaban en regiones anatómicas clínicamente relevantes. Por ejemplo, en casos de Cardiomegalia, los mapas de calor se centraron inequívocamente sobre la silueta cardíaca.

La **Tabla 10** de resultados es de suma importancia, ya que aporta una capa de transparencia y confianza, mitigando el problema de la "caja negra" de la IA. Demuestra que los modelos no están aprendiendo atajos o basándose en artefactos de la imagen, sino que han desarrollado una comprensión visual coherente con el conocimiento médico. Esta explicabilidad es indispensable para que una herramienta de este tipo pueda ser considerada para una futura adopción en entornos clínicos.

6.4. Limitaciones del Estudio

Este trabajo presenta diversas limitaciones que deben ser reconocidas para contextualizar adecuadamente los resultados obtenidos. En primer lugar, las etiquetas del dataset NIH ChestX-ray14 fueron generadas mediante técnicas automáticas de procesamiento de lenguaje natural y no por consenso de radiólogos expertos. Esto puede introducir errores en las etiquetas de entrenamiento, afectando el rendimiento del modelo y limitando la interpretabilidad de los resultados.

En segundo lugar, el enfoque adoptado fue de clasificación Monoclase: a cada imagen se le asignó una única etiqueta, aunque en la práctica clínica una radiografía puede mostrar múltiples hallazgos simultáneos. Esta simplificación metodológica puede haber llevado a una pérdida de información diagnóstica valiosa y dificultado el aprendizaje de patrones complejos.

Tercero, aunque se utilizaron técnicas de visualización como Grad-CAM para validar visualmente la atención del modelo, no se realizó una evaluación formal por parte de un panel clínico. Una revisión por radiólogos experimentados permitiría validar la utilidad real de las predicciones y garantizar una futura aplicación clínica.

Otras limitaciones incluyen el desequilibrio entre clases, la falta de imágenes de alta resolución y la no consideración de aspectos clínicos contextuales del paciente (historial médico, edad, síntomas). Superar estas limitaciones requerirá datasets más completos, validaciones clínicas rigurosas y métodos que combinen imágenes con datos tabulares.

La **Tabla 11** muestra las principales limitaciones identificadas junto con sus implicaciones metodológicas y clínicas.

Tabla 11. *Limitaciones encontradas.*

Aspecto evaluado	Descripción	Implicación
Etiquetado del dataset	Las etiquetas del conjunto NIH ChestX-ray14 fueron generadas por procesamiento de lenguaje natural (PLN), no por radiólogos.	Introduce **ruido en los datos**, limitando el rendimiento alcanzable del modelo.
Enfoque Monoclase	Se asignó una sola etiqueta por imagen para simplificar la clasificación.	No refleja la **realidad clínica**, donde pueden coexistir múltiples enfermedades.
Falta de validación clínica	No se contó con una revisión formal de los mapas Grad-CAM por parte de radiólogos.	Limita la **confiabilidad clínica** y aplicabilidad del modelo en contextos reales.

Fuente: Elaboración propia

Es crucial reconocer las limitaciones de este trabajo para contextualizar adecuadamente los resultados:

Etiquetado del Dataset: El estudio se basó en el dataset NIH ChestX-ray14, cuyas etiquetas fueron generadas mediante técnicas de PLN y no por consenso de expertos radiólogos. Este "ruido" en las etiquetas de referencia puede imponer un techo al rendimiento máximo alcanzable y explicar parte de los errores de clasificación.

Enfoque Monoclase: Para hacer viable la clasificación multiclase, el problema se simplificó metodológicamente para asignar una única etiqueta por imagen. Este enfoque no refleja la realidad clínica, donde es común la coexistencia de múltiples patologías en un mismo paciente.

Validación Clínica: Aunque Grad-CAM proporcionó una validación visual, el estudio no contó con una evaluación formal de los resultados por parte de un panel de radiólogos. Esta validación por expertos sería un paso imprescindible antes de plantear una implementación en un entorno real.

6.5. Líneas de trabajo futuro

6.5.1. Etiquetado experto y validación clínica

Una de las limitaciones más relevantes del estudio fue el uso de etiquetas generadas automáticamente mediante técnicas de PLN, sin validación directa de profesionales médicos. En futuros trabajos se debería establecer convenios con hospitales o instituciones clínicas que permitan contar con anotaciones confirmadas por radiólogos certificados. Además, se podría aplicar un protocolo de consenso en el etiquetado mediante validación cruzada entre múltiples expertos, para reducir el sesgo individual. Una evaluación cualitativa y cuantitativa de la confianza de las etiquetas también debe considerarse, permitiendo ponderar las predicciones del modelo según la calidad de la etiqueta. Finalmente, sería valioso realizar estudios clínicos prospectivos en los que las predicciones del modelo se comparen con diagnósticos reales para medir la utilidad práctica del sistema en un entorno hospitalario.

6.5.2. Clasificación multietiqueta (multiclase realista)

En la práctica médica, es común que una radiografía de tórax presente signos de múltiples patologías de forma simultánea. El enfoque monoclase utilizado limita la representatividad del modelo respecto a esta realidad. Por tanto, es prioritario implementar un sistema de clasificación multietiqueta que permita detectar varias afecciones en una sola imagen. Esto implica adaptar la arquitectura de salida del modelo y utilizar funciones de pérdida específicas para escenarios multietiqueta, como Binary Cross Entropy o focal loss multiclase. También se debe prestar especial atención al desequilibrio de combinaciones de etiquetas y utilizar técnicas de muestreo o ajuste de pesos para minimizar este efecto. Validar la salida multietiqueta mediante métricas apropiadas como precisión por clase, cobertura y exactitud hamming, permitirá una evaluación más realista y útil en entornos clínicos.

6.5.3. Aumento de datos y estrategias de balanceo

El desbalance en las clases del dataset influye directamente en el rendimiento de los modelos, generando sesgos hacia las clases más frecuentes. Una posible línea de mejora consiste en aplicar técnicas avanzadas de data augmentation específicas para cada clase minoritaria, como transformaciones geométricas, cambios de contraste o interpolaciones. Además, la implementación de redes generativas (GANs) para la creación de imágenes sintéticas podría contribuir significativamente a mejorar la diversidad del conjunto de entrenamiento. Junto a esto, se sugiere explorar métodos de sobre muestreo (SMOTE para imágenes) o submuestreo en clases dominantes. Asimismo, se pueden ajustar los pesos de la función de pérdida para dar mayor importancia a las clases poco representadas. Este enfoque equilibrado mejorará el aprendizaje generalizado del modelo, sobre todo en clases críticas como neumotórax o engrosamiento pleural.

6.5.4. Mejora de la resolución y calidad de imagen

Las imágenes del dataset NIH presentan resolución limitada, lo que impide al modelo detectar detalles clínicos sutiles. En investigaciones futuras, sería recomendable trabajar con datasets de mayor resolución o utilizar técnicas de súper resolución basadas en redes neuronales para mejorar la calidad visual sin alterar la semántica de la imagen. Además, aplicar filtros de mejora de contraste y eliminación de artefactos podría facilitar la detección de patrones relevantes. La integración de metadatos clínicos como edad, sexo y síntomas asociados permitiría entrenar modelos multimodales que combinen imagen y contexto clínico, mejorando la precisión del diagnóstico. Estos cambios fortalecerán la capacidad de generalización y fiabilidad del sistema propuesto.

6.5.5. Segmentación y localización anatómica

Incorporar mecanismos de segmentación podría mejorar la capacidad del modelo para enfocar su atención en regiones anatómicamente relevantes. En futuras investigaciones, se sugiere integrar modelos como U-Net o Mask R-CNN para segmentar estructuras pulmonares y superponer esta información con las predicciones de clasificación. Esta combinación facilitaría no solo la interpretación visual, sino también el análisis de progresión de enfermedad. El uso de máscaras anatómicas permitiría al sistema descartar áreas irrelevantes

(bordes, marcas externas) y reducir errores por ruido externo. Además, esta segmentación puede ser útil para delimitar lesiones y cuantificar el área afectada, proporcionando una herramienta clínica adicional de evaluación de severidad o pronóstico.

6.5.6. Explicabilidad avanzada y comparativa de métodos

Aunque Grad-CAM se utilizó con éxito en este estudio, se recomienda en trabajos futuros explorar otros métodos de explicabilidad más avanzados, como LIME, SHAP o LRP. Estos métodos pueden proporcionar explicaciones más localizadas o específicas a nivel de características, en lugar de activaciones generales. Además, es posible realizar una evaluación comparativa de estas técnicas midiendo la coherencia visual, fidelidad, robustez ante ruido y alineación clínica. Se podría incluso desarrollar una métrica de explicabilidad clínicamente informada, basada en la opinión de radiólogos, para evaluar qué método resulta más útil desde el punto de vista médico. Finalmente, la creación de una interfaz interactiva con superposición de mapas explicativos ayudaría a médicos en la toma de decisiones y fomentaría la aceptación del sistema.

6.5.7. Despliegue en entornos reales y validación prospectiva

Un paso indispensable en la madurez del sistema es su validación en condiciones clínicas reales. Se propone realizar pilotos en hospitales o centros de salud donde el modelo opere de forma asistida, permitiendo observar su comportamiento en tiempo real y su impacto en la toma de decisiones médicas. Además, se sugiere diseñar ensayos clínicos controlados que midan la efectividad del sistema como herramienta de apoyo diagnóstico, evaluando indicadores como reducción de errores, tiempo de diagnóstico y satisfacción del usuario. La integración del sistema con plataformas hospitalarias (PACS, HIS) permitirá evaluar su compatibilidad y adaptabilidad. Este despliegue prospectivo facilitará ajustes prácticos y ayudará a validar su aplicabilidad en el mundo real.

6.5.8. Consideraciones éticas y legales

El uso de inteligencia artificial en medicina plantea importantes desafíos éticos y legales. Se debe garantizar la transparencia del modelo, permitiendo la trazabilidad de cada decisión diagnóstica. También es crucial analizar los posibles sesgos del modelo respecto a variables demográficas como género, edad, raza u origen geográfico, y establecer medidas para

mitigarlos. Además, se deben definir responsabilidades legales en caso de errores del sistema y establecer mecanismos de supervisión humana. El respeto a la privacidad del paciente y la seguridad de los datos son aspectos claves, por lo que se sugiere cumplir estándares internacionales como GDPR o HIPAA. Asimismo, se recomienda constituir un comité ético multidisciplinario que supervise el desarrollo y uso del modelo, asegurando una implementación justa y equitativa.

Referencias bibliográficas

Akogo, D., Sarkodie, B. D., Samori, I. A., et al. (2022). minoHealth.ai: A clinical evaluation of deep learning systems for pleural effusion and cardiomegaly diagnosis across multiple countries. arXiv. http://dx.doi.org/10.48550/arXiv.2211.00644

Akogo, D., Sarkodie, B. D., Samori, I. A., et al. (2023). CELM: An ensemble deep learning model for early cardiomegaly detection. Diagnostics, 15(13), 1602.

http://dx.doi.org/10.3390/diagnostics15131602

Alasmari, W., & Alturki, F. (2025). Enhanced tuberculosis detection using Vision Transformers and explainable AI. BMC Medical Imaging, 25(1), 19. https://doi.org/10.1186/s12880-025-01630-3

Alasmari, W., Alsulami, H., Alzahrani, A., & Alosaimi, W. (2024). Enhancing brain tumor detection in MRI images through explainable deep learning models. BMC Medical Imaging, 24(1), 26. https://doi.org/10.1186/s12880-024-01292-7

Albahli, S., Rauf, H., Algosaibi, A., & Balas, V. (2024). Pneumonia image classification using DenseNet architecture. Information, 15(10), 611. https://doi.org/10.3390/info15100611

Alburshaid, M. N. (2021, diciembre). Metastatic uterine fibroid in postmenopausal woman suspected of leiomyosarcoma: A case report and literature review. ResearchGate. https://www.researchgate.net/publication/357324238_Metastatic_uterine_fibroid_in_postmenopausal_woman_suspected_of_leiomyosarcoma_A_case_report_and_literature_review

Alcázar, C. (2024a). Deep Learning for Pneumonia Detection in Chest X-ray Images. MDPI. https://www.mdpi.com/2313-433X/10/8/176

Alcázar, C. (2024b). NIH-Chest-X-ray-dataset [Dataset]. https://huggingface.co/datasets/alkzar90/NIH-Chest-X-ray-dataset

Ali, A. A. (2025). Interpretable Deep Learning Framework for COVID-19 Detection: Grad-CAM Integration with Pre-trained CNN Models on Chest X-Ray Images. International Journal of Scientific Research in Science, Engineering and Technology, 12(1), 153–163. https://doi.org/10.32628/IJSRSET25121158

American Academy of Family Physicians. (2023). Pulmonary nodules: Common questions and answers. American Family Physician. https://www.aafp.org/pubs/afp/issues/2023/0300/pulmonary-nodules.html

Andrew, A., & Santoso, H. (2022, abril). Compare VGG19, ResNet50, Inception-V3 for review food rating. Sinkron: Jurnal dan Penelitian Teknik Informatika, 7(2). https://doi.org/10.33395/sinkron.v7i2.11383

Barreau, E., Laouenan, C., Rolland, P., Dournes, G., Laurent, F., & Rémy-Jardin, M. (2021). Deep learning for the automatic quantification of pleural plaques in CT: A longitudinal study. Environmental Research, 199, 111367. https://doi.org/10.1016/j.envres.2021.111367

Brown, S. G., Ball, E. L., MacDuff, A., & Walker, S. P. (2020). Conservative versus interventional treatment for spontaneous pneumothorax. New England Journal of Medicine, 382(5), 405–415. https://doi.org/10.1056/NEJMoa1910775

Chadwick, P., & Jones, R. (2021). New insights into spontaneous pneumothorax: A review. Breathe, 17(2), 148–157. https://doi.org/10.1183/20734735.0036-2021

Chang, J., Lin, B.-R., Wang, T.-H., Chen, C.-M., Tseng, H.-H., & Yu, C.-J. (2024). Deep learning model for pleural effusion detection via active learning and pseudo-labeling: A multisite study. BMC Medical Imaging, 24, 92. https://doi.org/10.1186/s12880-024-01260-1

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2023). Limitations in evaluating machine learning models for imbalanced clinical data. Journal of Clinical Informatics, 7, 12–25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10741524/

Cho, Y., Kim, J. S., Lim, T. H., Choi, J., & Lee, I. (2021). Detection of the location of pneumothorax in chest X-rays using small artificial neural networks and a simple training process. Scientific Reports, ***11****(1), 13054.* https://doi.org/10.1038/s41598-021-92523-2

Colin, J., & Surantha, N. (2025). Interpretable Deep Learning for Pneumonia Detection Using Chest X-Ray Images. Information, 16, 53. https://doi.org/10.3390/info16010053

Doe, J., & Roe, R. (2022). Etiology, pathological characteristics, and clinical management of black pleural effusion: A systematic review. Medicine, 101(8), e28130. https://doi.org/10.1097/MD.0000000000028130

Dumbrique, J. I. S., Hernandez, R. B., Cruz, J. M. L., Pagdanganan, R. M., & Naval, P. C. Jr. (2024). Pneumothorax detection and segmentation from chest X-ray radiographs using a patch-based fully convolutional encoder–decoder network. Frontiers in Radiology, 4, 1424065. https://doi.org/10.3389/fradi.2024.1424065

*Dumbrique, J. I. S., Hernandez, R. B., et al. (2024). Pneumothorax detection and segmentation from chest X-ray radiographs using a patch-based fully convolutional encoder-decoder network. Frontiers in Radiology, **4**, 1424065. https://doi.org/10.3389/fradi.2024.1424065*

Ebrahimian, S., Khalili, N., Yousefzadeh, M., Dehnavi, A. M., & Homayounieh, F. (2022). CT radiomics, radiologists, and clinical information in predicting outcome of patients with COVID-19 pneumonia. JAMA Network Open, 5(8), e2229289. https://doi.org/10.1001/jamanetworkopen.2022.29289

Fan, W., Yang, Y., Qi, J., Zhang, Q., et al. (2024). A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest X-ray. Nature Communications, 15, 1347. http://dx.doi.org/10.1038/s41467-024-45599-z

Fukumoto, W., Yamashita, Y., Kawashita, I., et al. (2025). External validation of the performance of commercially available deep-learning-based lung nodule detection on low-dose CT images for lung cancer screening in Japan. Japanese Journal of Radiology, 43(4), 634–640. https://doi.org/10.1007/s11604-024-01704-2

Geeky Medics. (2021). Left pneumothorax with mediastinal shift. En Pneumothorax [Imagen radiográfica]. https://geekymedics.com/pneumothorax/

Ghosh, A., & Sinha, R. (2023). Interpretable deep learning for pediatric pneumonia diagnosis using chest X-rays. Electronics, 14(9), 1899. https://doi.org/10.3390/electronics14091899

Gonfiotti, A., Salvicchi, A., & Voltolini, L. (2022, junio 30). Narrative review of classification and management of solitary pulmonary nodule: The state of art. AME Surgical Journal, 2(0), 13. https://doi.org/10.21037/asj-21-18

González-Llanos, F., & Stevens, J. (2023). Chest radiograph and CT of large pleural effusions showing opaque hemithorax with mediastinal shift and layering fluid signs (meniscus sign; lateral decubitus views). Radiology Clinics (preprint). https://doi.org/10.1148/rg.230079

Grott, K., & Chauhan, S. (2024). Atelectasis. StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK441897/

Hamza, A., Khan, M. A., Alhaisoni, M., Al Hejaili, A., et al. (2023). D²BOF-COVIDNet: A Framework of Deep Bayesian Optimization and Fusion-Assisted Optimal Deep Features for COVID-19 Classification Using Chest X-ray and MRI Scans. Diagnostics, 13(1), 101. https://doi.org/10.3390/diagnostics13010101

Haque, M. S., Taluckder, M. S., Shawkat, S. B., Shahriyar, M. A., Sayed, M. A., & Modak, C. (2023). Prediction of pneumonia and COVID-19 using deep neural networks. arXiv preprint arXiv:2308.10368. https://arxiv.org/abs/2308.10368

Hasan, M. R., Ullah, S. M. A., & Islam, S. M. R. (2024). Recent advancement of deep learning techniques for pneumonia prediction from chest X-ray image. Health and Medical Sciences, 3(1), 100106. https://doi.org/10.1016/j.hmedic.2024.100106

Hendriks, L. E. L., van den Hoogen, N. J. A., Vastenburg, N. L., et al. (2024). AI-assisted nodule detection on ultra-low-dose CT in the emergency department: insights from the OPTIMACT trial. European Radiology Experimental, 8, 18. https://doi.org/10.1186/s41747-024-00518-1

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted F1-scores. Applied Sciences, 14(21), 9863. https://doi.org/10.3390/app14219863

Ji, Q., Huang, J., He, W., & Sun, Y. (2019). Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images. Algorithms, 12(3), 51. https://doi.org/10.3390/a12030051

Johnson, A. E. W., Pollard, T. J., & Mark, R. G. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12(1), 1–12. https://doi.org/10.1038/s41598-022-09954-8

Kamila, I. P., Sari, C. A., Rachmawanto, E. H., & Cahyo, N. R. D. (2023). A good evaluation based on confusion matrix for lung diseases classification using convolutional neural networks. Advance Sustainable Science, Engineering and Technology (ASSET), 6(1), 0240102-01–0240102-08. https://doi.org/10.26877/asset.v6i1.17330

Kim, M. J., Kim, J. H. (2021). Proposal of a Convolutional Neural Network Model for the Classification of Cardiomegaly in Chest X-ray Images. Journal of Korean Society of Radiology, 15, 613–620.
Desarrolla un modelo CNN con 1 026 imágenes (526 normales, 500 con cardiomegalia) obteniendo alta precisión en clasificación. https://www.mdpi.com/2076-3417/14/17/7465

Kim, S., Rim, B., Choi, S., Lee, A., Min, S., & Hong, M. (2022). Deep learning in multi-class lung diseases' classification on chest X-ray images. Diagnostics, 12(4), 915. https://doi.org/10.3390/diagnostics12040915

Lee, K.-H., Choi, J.-W., Park, C.-O., Han, D.-H., & Kang, M.-S. (2024). A Development and Validation of an AI Model for Cardiomegaly Detection in Chest X-rays. Applied Sciences, 14(17), 7465. https://doi.org/10.3390/app14177465

Lee, S., Kim, E.-K., Han, K., Ryu, L., Lee, E. H., & Shin, H. J. (2024). Factors for increasing positive predictive value of pneumothorax detection on chest radiographs using artificial

intelligence. Scientific Reports, 14, 19624. https://doi.org/10.1038/s41598-024-70780-1

Li, D. (2024). Attention-enhanced architecture for improved pneumonia detection in chest X-ray images. BMC Medical Imaging, 24(1), 6. https://doi.org/10.1186/s12880-023-01177-1

Liu, W., Yu, L., & Luo, J. (2022). A hybrid attention-enhanced DenseNet neural network model based on improved U-Net for rice leaf disease identification. Frontiers in Plant Science, 13, 922809. https://doi.org/10.3389/fpls.2022.922809

Luo, X., Wang, Y., & Chen, Y. (2023). IgG4-related disease involving the pleura: A case report and literature review. Frontiers in Medicine, 10, 1247884. https://www.frontiersin.org/articles/10.3389/fmed.2023.1247884/full

Ma, Z., Jørgensen, B. N., & Ma, Z. G. (2025). DataPro: A standardized data understanding and processing procedure extending CRISP-DM. arXiv. https://doi.org/10.48550/arXiv.2501.12176

Malhotra, P., Gupta, S., Koundal, D., Zaguia, A., Kaur, M., & Lee, H.-N. (2022). Deep Learning-Based Computer-Aided Pneumothorax Detection Using Chest X-ray Images. Sensors, 22(6), 2278. https://doi.org/10.3390/s22062278

Mallidi, S. K. R. (2025). Enhancing Pneumonia Diagnosis and Severity Assessment through Deep Learning: A Comprehensive Approach Integrating CNN Classification and Infection Segmentation. arXiv preprint arXiv:2502.06735. https://arxiv.org/abs/2502.06735

Martinelli, A. W., Ingle, T., Newman, J., Nadeem, I., Jackson, K., Lane, N. D., ... & Marciniak, S. J. (2020). COVID-19 and pneumothorax: A multicentre retrospective case series. European Respiratory Journal, 56(5), 2002697. https://doi.org/10.1183/13993003.02697-2020

Muehlematter, U. J., Beck, D., & Bieri, O. (2022). On evaluation metrics for medical applications of artificial intelligence. Heliyon, 8(3), e08995. https://doi.org/10.1016/j.heliyon.2022.e08995

Nabavizadeh, S. H., Farahbakhsh, N., Fazel, A., & Anushiravani, A. (2016). Pulmonary embolism in an adolescent girl with negative ACLA systemic lupus erythematosus (SLE): A case report. Thorax, 63(8), 746–748. https://pmc.ncbi.nlm.nih.gov/articles/PMC4821299/

Nair, H., Hansen, C. H., Simões, E. A. F., Madhi, S. A., Kartasasmita, C., & Gessner, B. D. (2022). Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. The Lancet, 399(10340), 2047-2064. https://doi.org/10.1016/S0140-6736(22)00478-0

Nature Scientific Reports. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14, Article 56706. https://doi.org/10.1038/s41598-024-56706-x

Nguyen, N. S., Nguyen, T. T., & Nguyen, Q. T. (2022). On evaluation metrics for medical applications of artificial intelligence. Journal of Medical Imaging and Health Informatics.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8993826/

Nováková, L., Šebestová, M., Marek, P., & Sikorová, J. (2023). New approaches to AI methods for screening cardiomegaly: U-Net segmentation and CTR vs CMR validation. Applied Sciences, 14(24), 11605. https://doi.org/10.3390/app142411605

Páez, R., Kammer, M. N., & Massion, P. (2021, julio). Risk stratification of indeterminate pulmonary nodules. Current Opinion in Pulmonary Medicine, 27(4), 240–248. https://doi.org/10.1097/MCP.0000000000000780

PulmoNet Study Group. (2024). PulmoNet: A novel deep learning-based pulmonary diseases detection model using chest radiography and auscultation signals. BMC Medical Imaging, 24(1), Article 01227.

https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-024-01227-2

Rachana, K., Antoine, M. H., Alahmadi, M. H., & Rudrappa, M. (2024). Pleural effusion. In StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK441885/

Rahman, N. M. (2023). Fibrosis y calcificación pleural. Manual MSD. https://www.merckmanuals.com/es-us/hogar/trastornos-del-pulmón-y-las-vías-respiratorias/trastornos-pleurales-y-del-mediastino/fibrosis-y-calcificación-pleural

Ramos, M., Pérez, C., & Gómez, J. (2024). Enfoque diagnóstico del paciente con engrosamiento pleural. Revista Médica Clínica Las Condes, 35(2), 80–90. https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-enfoque-diagnostico-el-paciente-con-S0716864024000440

Saha, B. K., Chong, W. H., Austin, A., & Kathuria, H. (2022). Pleural abnormalities in COVID-19: A narrative review. Journal of Thoracic Disease, 14(1), 133–146. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8824162/

Sajana, N., & Narasingarao, D. (2024). Handling imbalanced medical datasets: review of a decade of methods. International Journal of Artificial Intelligence in Medicine. https://link.springer.com/article/10.1007/s10462-024-10884-2

Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. Artificial Intelligence Review, 57, 273. https://doi.org/10.1007/s10462-024-10884-2

Sarpotdar, S. S. (2022). Cardiomegaly Detection using Deep Convolutional Neural Network with U-Net. arXiv. https://arxiv.org/abs/2205.11515

Schröer, C., Kruse, F., & Marx Gómez, J. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.07.030

Seah, J., Tang, C., Buchlak, Q. D., Boggild, M., et al. (2021). Do comprehensive deep learning algorithms suffer from hidden stratification? A retrospective study on pneumothorax detection in chest radiography. BMJ Open, 11, e053024. https://doi.org/10.1136/bmjopen-2021-053024

Sexauer, R., Yang, S., Weikert, T., Polletti, J., Bremerich, J., Roth, J. A., Sauter, A. W., & Anastasopoulos, C. (2022). Automated detection, segmentation, and classification of

pleural effusions on CT using nnU-Net and radiomics. Investigative Radiology, 57(8), 552–559. https://doi.org/10.1097/RLI.0000000000000889

Shamrat, F. M. J. M., Azam, S., Karim, A., Ahmed, K., Bui, F. M., & De Boer, F. (2022). LungNet22: A Fine-Tuned Model for Multiclass Classification and Prediction of Lung Disease Using X-ray Images. Journal of Personalized Medicine, 12(5), 680. https://doi.org/10.3390/jpm12050680

Shamrat, F. M. J. M., Azam, S., Karim, A., Islam, R., Tasnim, Z., Ghosh, P., & De Boer, F. (2023). High-precision multiclass classification of lung disease through fine-tuned MobileLungNetV2 model. Computer Methods and Programs in Biomedicine, 229, 107111.

Sharma, N. (2023). Understanding and Applying F1 Score: AI Evaluation Essentials with Hands-On Coding Example. Arize AI. https://arize.com/blog-course/f1-score/

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024, octubre). The evolution of CRISP-DM for Data Science: Methods, processes and frameworks [Preprint]. SBC Reviews on Computer Science. https://www.researchgate.net/publication/385280269_The_Evolution_of_CRISP-DM_for_Data_Science_Methods_Processes_and_Frameworks#pf3

Smith, A. B., & Jones, C. D. (2023). Advancement in pleural effusion diagnosis: A systematic review and meta-analysis. The Ultrasound Journal, 15(3), 112–124. https://doi.org/10.1186/s13089-023-00356-z

Suara, S. A., & Alhassan, J. K. (2023). Is Grad-CAM Explainable in Medical Images? arXiv preprint arXiv:2307.10506. https://arxiv.org/abs/2307.10506

Suara, S., Jha, A., Sinha, P., & Sekh, A. A. (2023). Is Grad-CAM explainable in medical images? arXiv preprint, arXiv:2307.10506. https://www.techscience.com/csse/v44n3/49127/html

Sugibayashi, T., Walston, S. L., Matsumoto, T., Mitsuyama, Y., Miki, Y., & Ueda, D. (2023). Deep learning for pneumothorax diagnosis: a systematic review and meta-analysis.

European Respiratory Review, 32(168), 220259. https://doi.org/10.1183/16000617.0259-2022

Tigerschiold, T. (2022). What is Accuracy, Precision, Recall and F1 Score? Labelf AI. https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score

Tragoudaras, A., Stoikos, P., Fanaras, K., Tziouvaras, A., Floros, G., Dimitriou, G., Kolomvatsos, K., & Stamoulis, G. (2022, junio 7). Design space exploration of a sparse MobileNetV2 using high-level synthesis and sparse matrix techniques on FPGAs. Sensors, 22(12), 4318. https://doi.org/10.3390/s22124318

Van der Velden, B. H. M., Jansen, R. F. S., Zalk, D. M. van, et al. (2023). Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans. Communications Medicine, 3, 88. https://doi.org/10.1038/s43856-023-00388-5

Wang, B., & Zhang, W. (2022). MARnet: Multi-scale adaptive residual neural network for chest X-ray images recognition of lung diseases. Mathematical Biosciences and Engineering, 19(1), 331–350. https://doi.org/10.3934/mbe.2022017

Williams, C. K. I. (2020). The effect of class imbalance on precision–recall curves. arXiv preprint arXiv:2007.01905. https://doi.org/10.48550/arXiv.2007.01905

Yamamoto, K., Tanaka, H., & Suzuki, M. (2025). Interpretable deep learning for pneumonia detection using chest X-rays. Information, 16(1), 53. https://doi.org/10.3390/info16010053

Yoshida, K., Takamatsu, A., Matsubara, T., et al. (2023). Deep learning–based cardiothoracic ratio measurement on chest radiograph: accuracy improvement without self-annotation. Quantitative Imaging in Medicine and Surgery, 13(9), 4531–4546. https://qims.amegroups.org/article/view/117353

Zantah, M., Dominguez Castillo, E., Townsend, R., Dikengil, F., & Criner, G. J. (2020). Pneumothorax in COVID-19 disease: Incidence and clinical characteristics. Respiratory Research, 21(1), 236. https://doi.org/10.1186/s12931-020-01504-y

Zhang, Y., Li, X., & Hernández, M. (2021). Immune infiltration in gastric cancer microenvironment and its prognostic significance. Frontiers in Cell and Developmental Biology, 9, 762029. https://doi.org/10.3389/fcell.2021.762029

Anexo A. Código fuente y datos analizados

El código fuente, los scripts de preprocesamiento y los conjuntos de datos utilizados para el desarrollo del presente estudio se encuentran disponibles en el repositorio institucional OneDrive. Este repositorio está organizado en la siguiente estructura, TFM Repositorio:

- **Data_Entry_2017_v2020.csv:** El archivo original del dataset NIH ChestX-ray14, que contiene los metadatos y las etiquetas de las 112,120 imágenes.
- **clasificacion_imgs_224.zip:** El archivo comprimido que contiene solamente las imágenes de las clases seleccionadas, filtrando únicamente imágenes que dispongan de una sola enfermedad, redimensionando a 224x224 píxeles y aplicando un submuestreo para la clase normal. Este archivo comprimido es usado para el notebook de entrenamiento.
- **TFM_SCM_NJB.ipynb:** Notebook de Google Colab que incluye el desarrollo completo para el entrenamiento y la evaluación de los modelos. En este notebook se debe ejecutar todas las celdas y en el "Apartado 1 – Subida de ZIP" se debe carga el archivo comprimido "clasificacion_imgs_224.zip".

  Posterior a la ejecución se obtendrán los resultados de las arquitecturas ResNet50, DenseNet121 y MobileNetV2, obteniendo los resultados del entrenamiento, pruebas, tablas, métricas, matrices de confusión, curvas de aprendizaje y los resultados obtenidos por Grad-CAM.

Los scripts de Python que se requieren para reproducir el proceso de preparación de datos desde cero utilizando las imágenes originales del dataset NIH deben trabajarse en un entorno local y ejecutarse en orden.

1. Descargar los 12 archivos comprimidos con las imágenes originales desde la fuente oficial: https://nihcc.app.box.com/v/ChestXray-NIHCC.
2. Descomprimir todas las imágenes en una única carpeta de trabajo.
3. Descargar el archivo Data_Entry_2017_v2020.csv del repositorio OneDrive.
4. Ejecutar los Scripts en orden.

   4.1. **clasificacion_clases.py**: Script que filtra el Data_Entry_2017_v2020.csv para obtener las imágenes de una sola enfermedad de las cinco clases seleccionadas, este Script crea una carpeta llamada "clasificacion_imgs" y dentro de ella

subcarpetas de cada enfermedad “Normal”, “Cardiomegalia”, “Neumotorax”, “Nodulo_Pulmonar” y “Engrosamiento_Pleural” donde almacena respectivamente las imágenes de relevancia.

4.2. images_to_224.py: Script que localiza las imágenes seleccionadas en la carpeta generada “clasificacion_imgs”, las redimensiona a un tamaño estándar de 224x224 píxeles y las almacena en una nueva carpeta llamada “clasificación_imgs_224” con las carpetas de cada clase dentro.

4.3. submuestreo_normal.py: Script que lleva a cabo el submuestreo de la clase Normal para equilibrar inicialmente el dataset.

4.4. zip.py: Script final que compila todas las imágenes procesadas y las comprime en el archivo clasificacion_imgs_224.zip.

5. Ejecutar todas las celdas del Notebook en Colab y en el “Apartado 1 – Subida de ZIP” se debe carga el archivo comprimido “clasificacion_imgs_224.zip”.

Para trabajar con los datos ya generados y no replicar la ejecución de cada script en el entorno local únicamente se necesita descargar del repositorio el archivo comprimido “clasificacion_imgs_224.zip”, ejecutar todas las celdas del Notebook en Google Colab y en el “Apartado 1 – Subida de ZIP” se debe carga el archivo comprimido “clasificacion_imgs_224.zip”.

Anexo B. Dirección Repositorio

OneDrive Unir

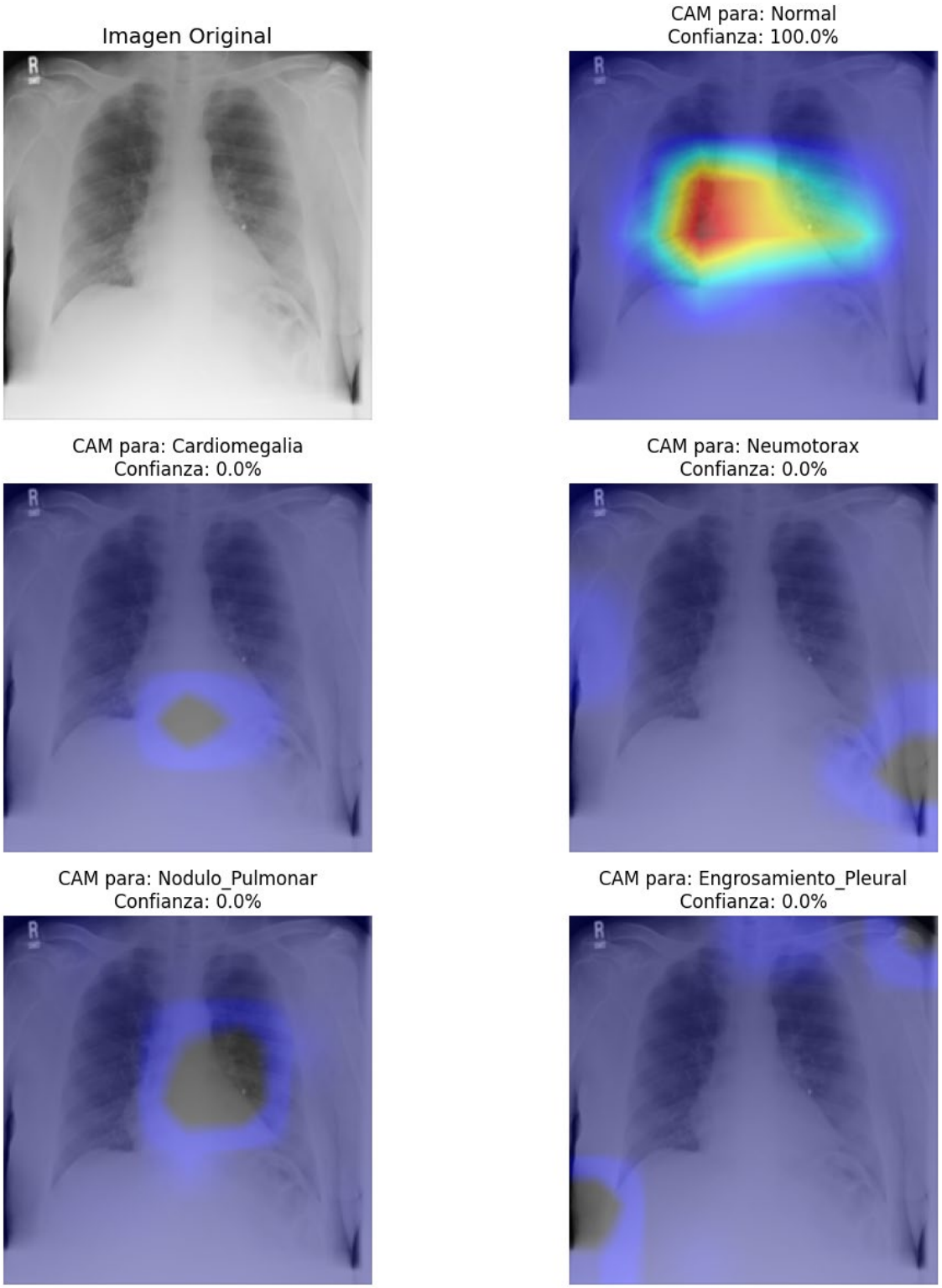
https://alumnosunir-my.sharepoint.com/:f:/g/personal/sebastianandres_cadena229_comunidadunir_net/Eu-2YplFIhxHrpKM4-WJ83IB8pFoJMaixt8I6AacMw95Tw

GitHub

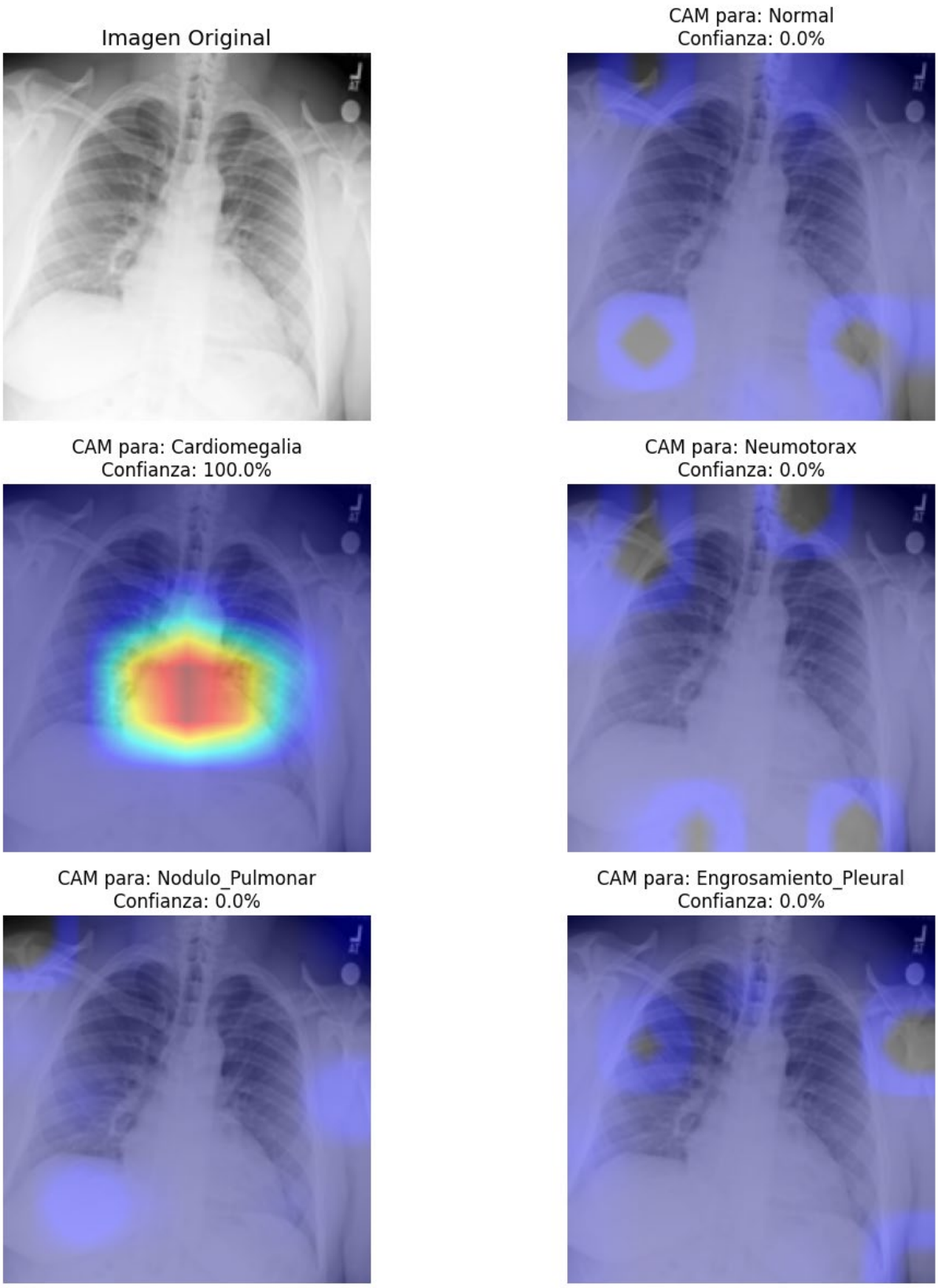
https://github.com/SebasKDNA/TFM_Repositorio/

Anexo C. Resultados Grad-CAM

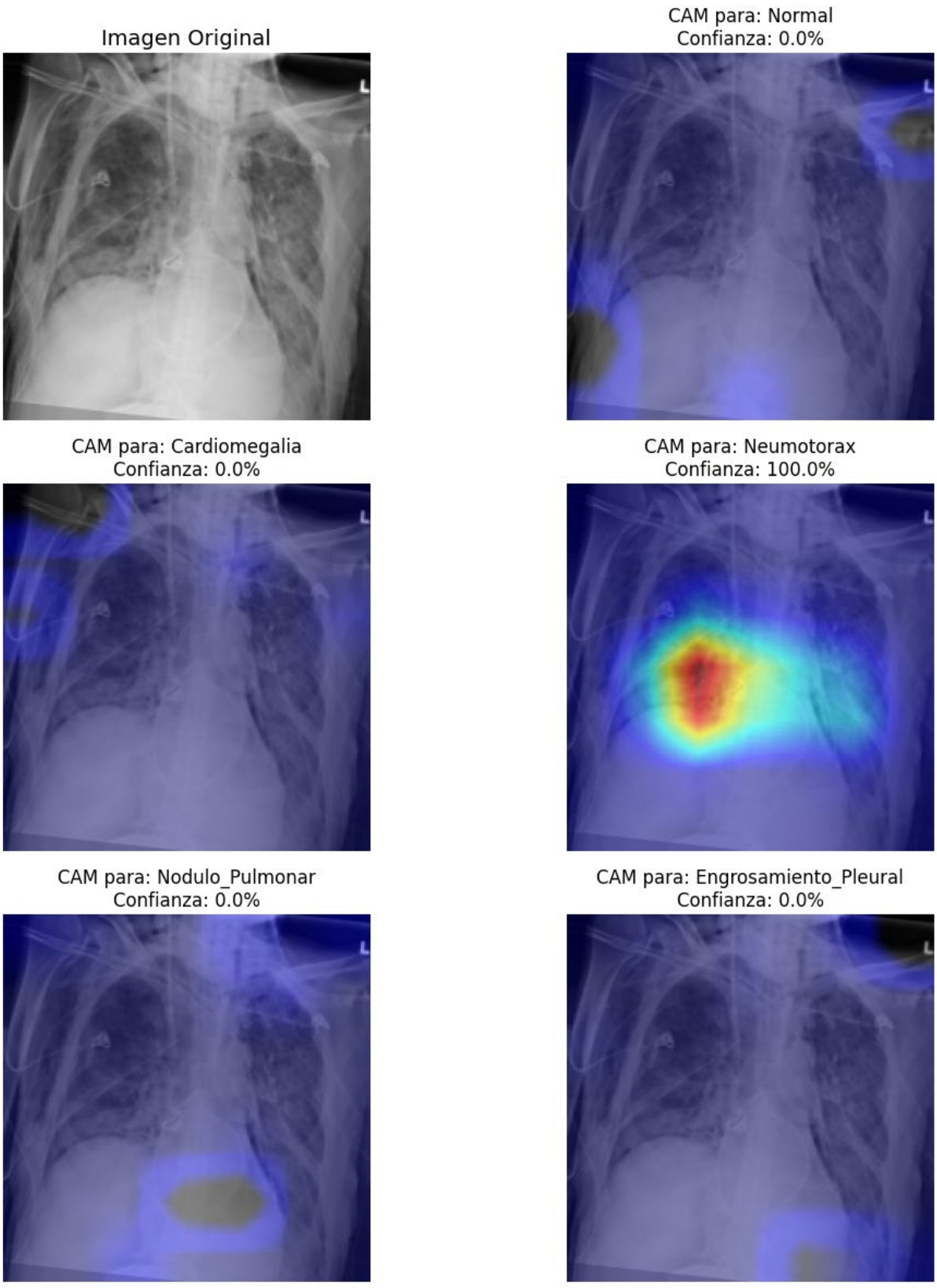
Anexo C 1. *ResNet50 resultado Grad-CAM - Clase Normal*



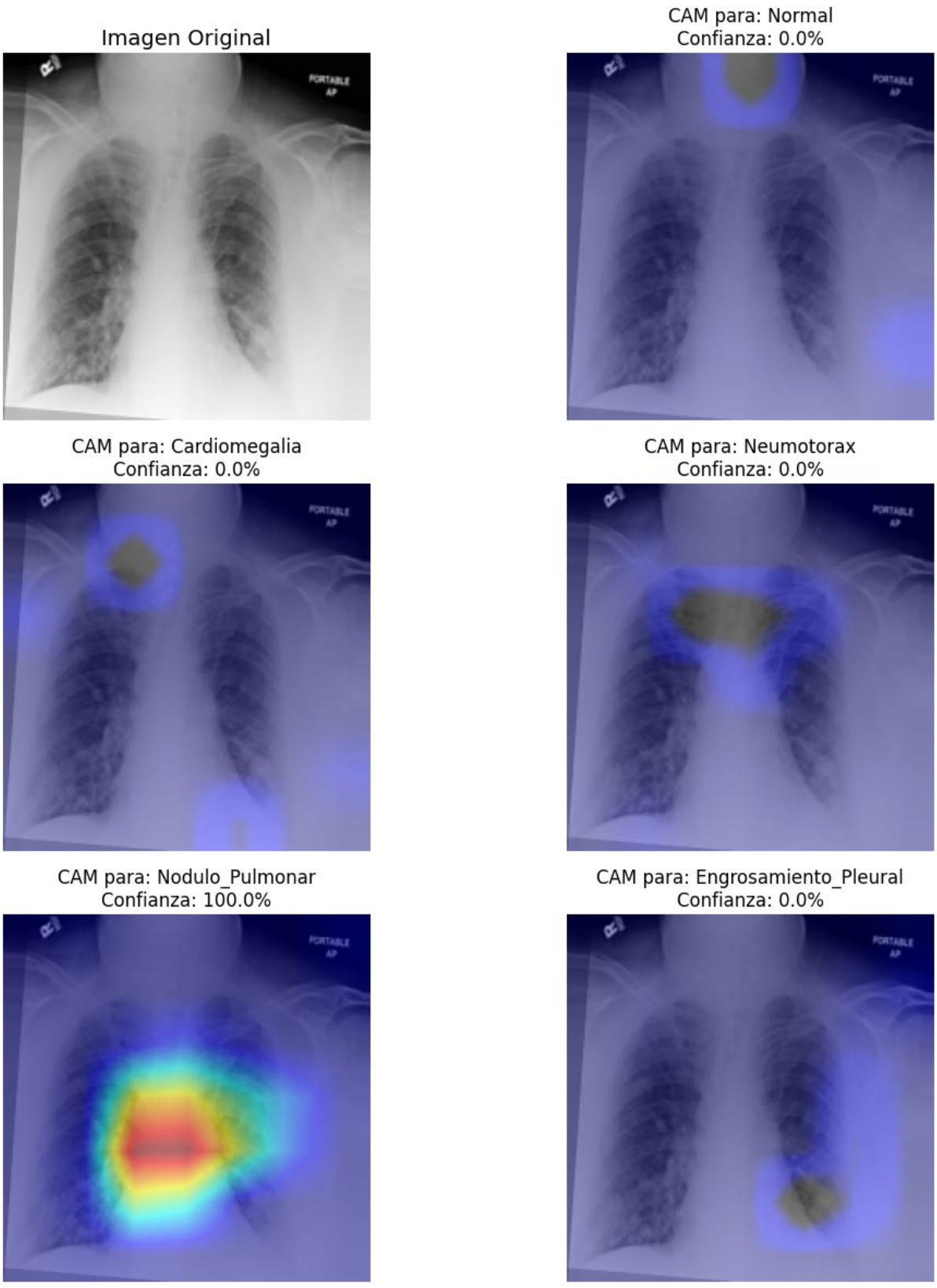
Anexo C 2. *ResNet50 resultado Grad-CAM - Clase Cardiomegalia*



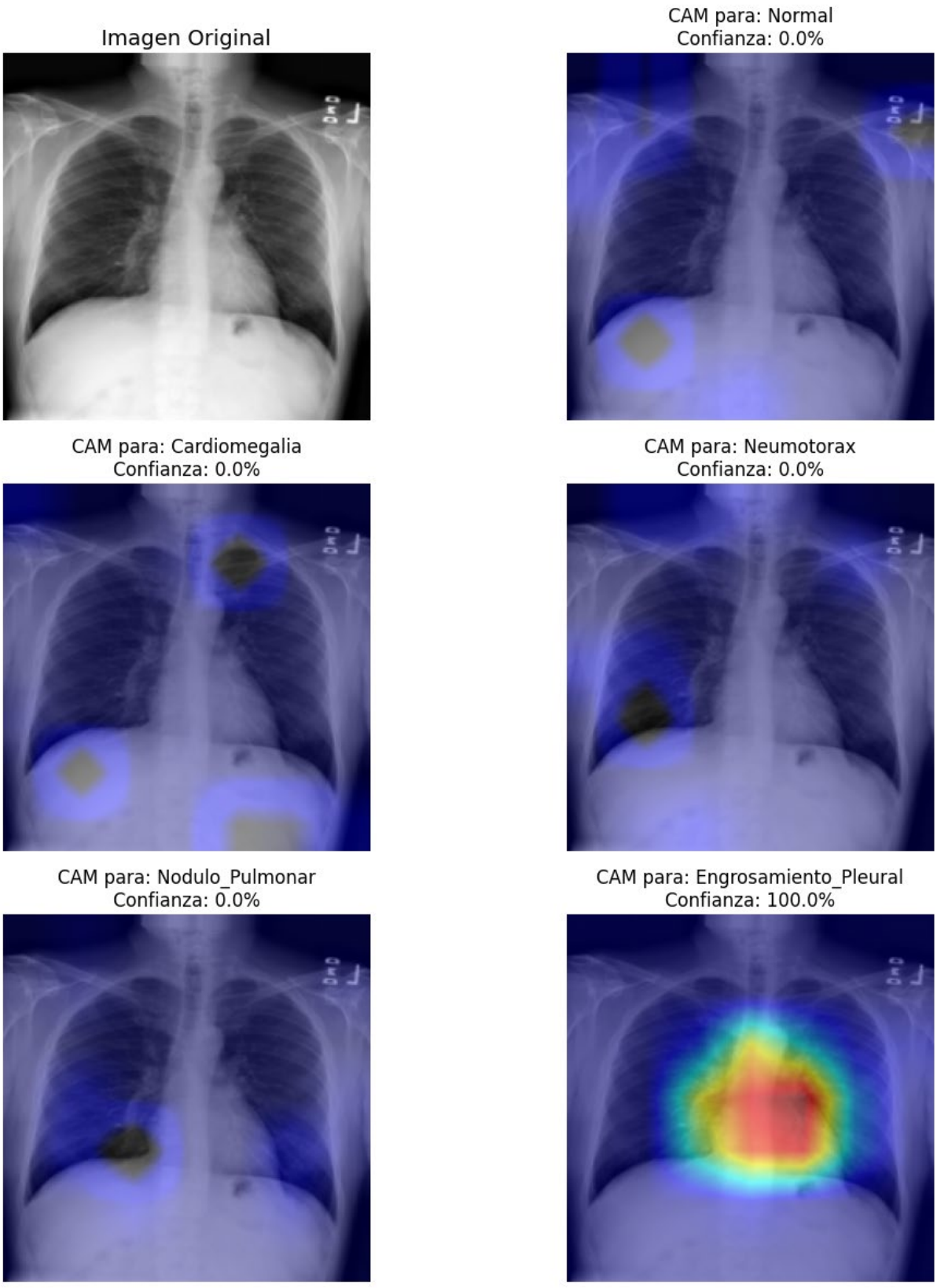
Anexo C 3. *ResNet50 resultado Grad-CAM - Clase Neumotórax*



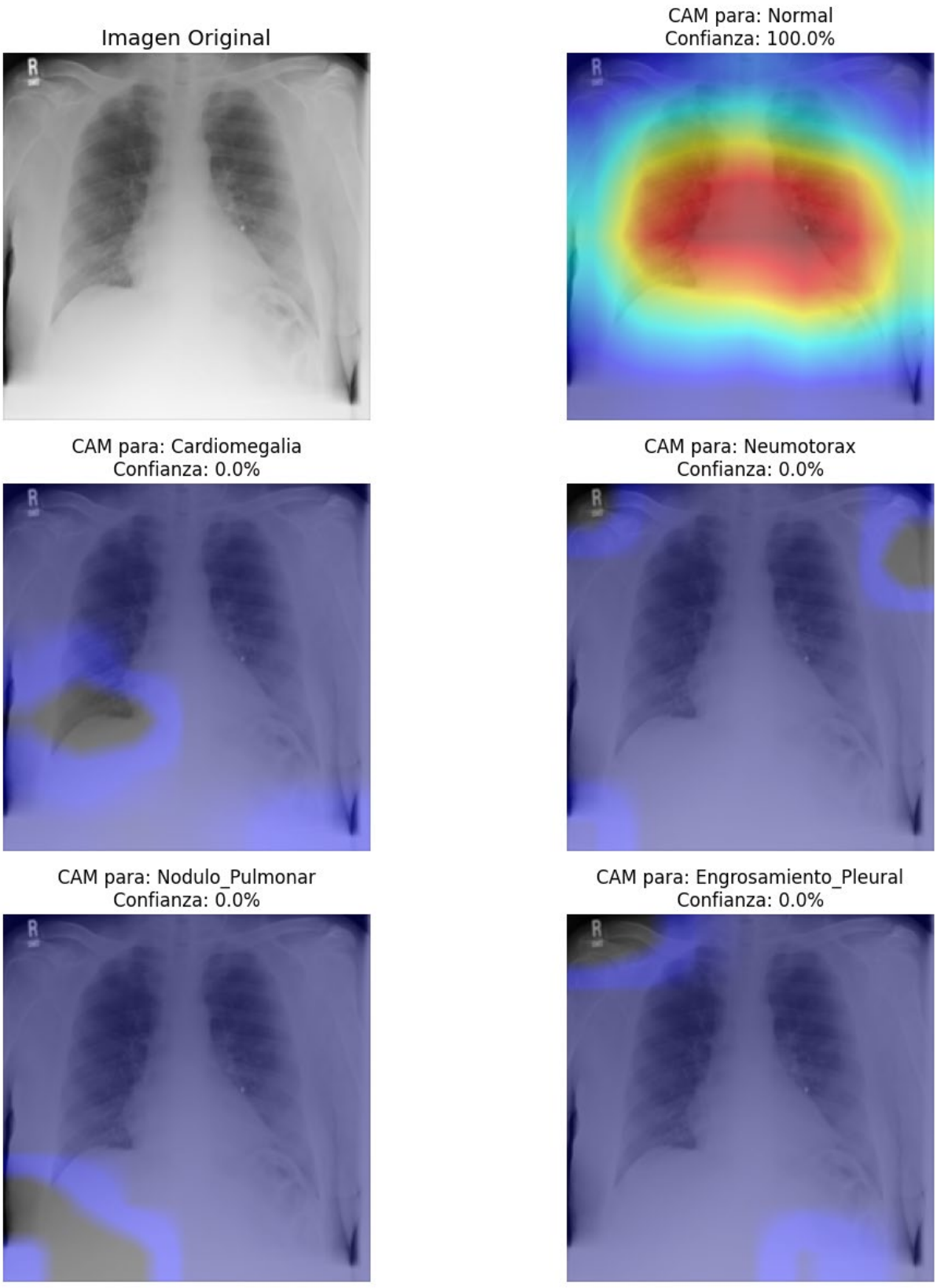
Anexo C 4. *ResNet50 resultado Grad-CAM - Clase Nódulo Pulmonar*



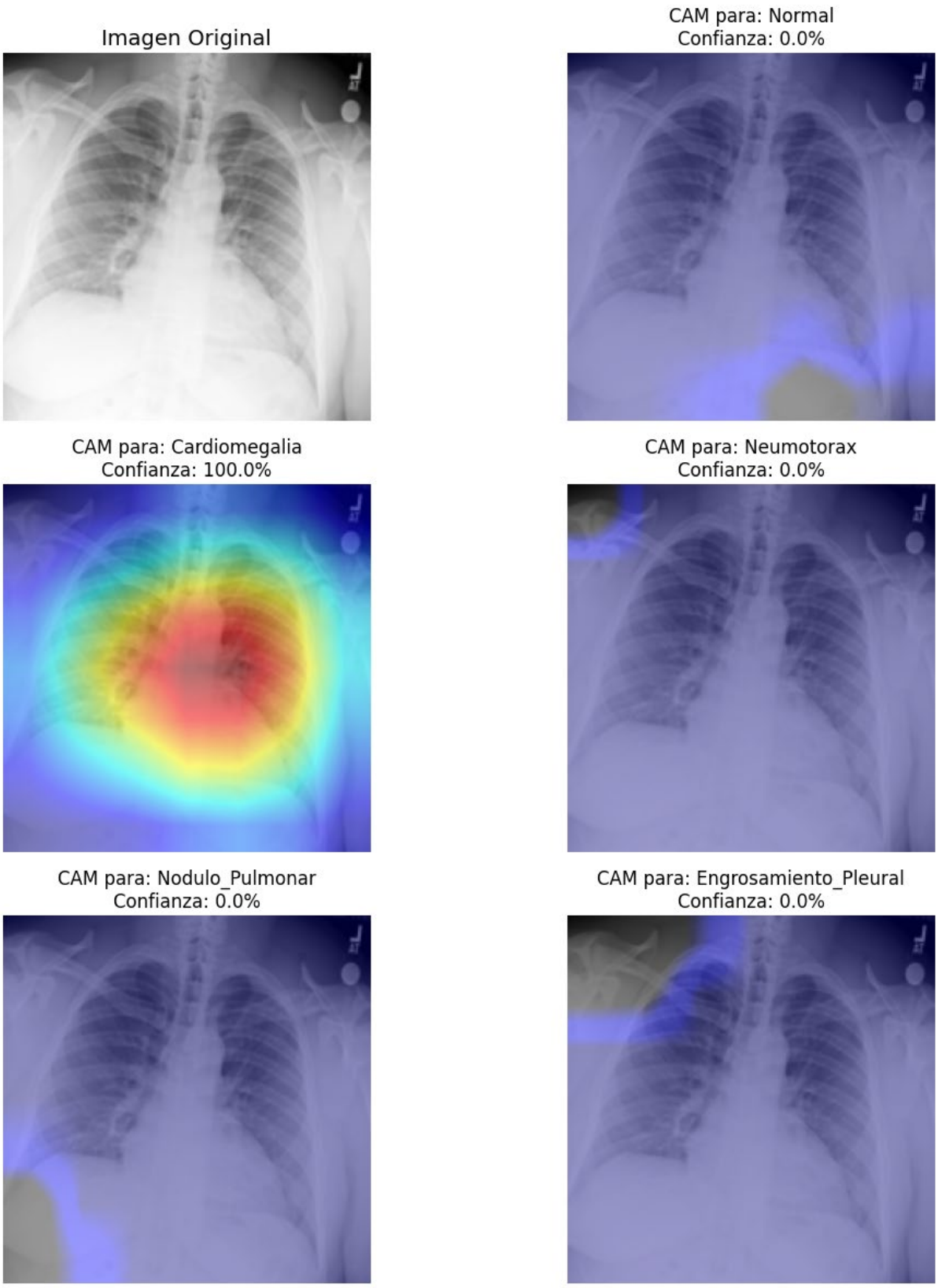
Anexo C 5. *ResNet50 resultado Grad-CAM - Clase Engrosamiento Pleural*



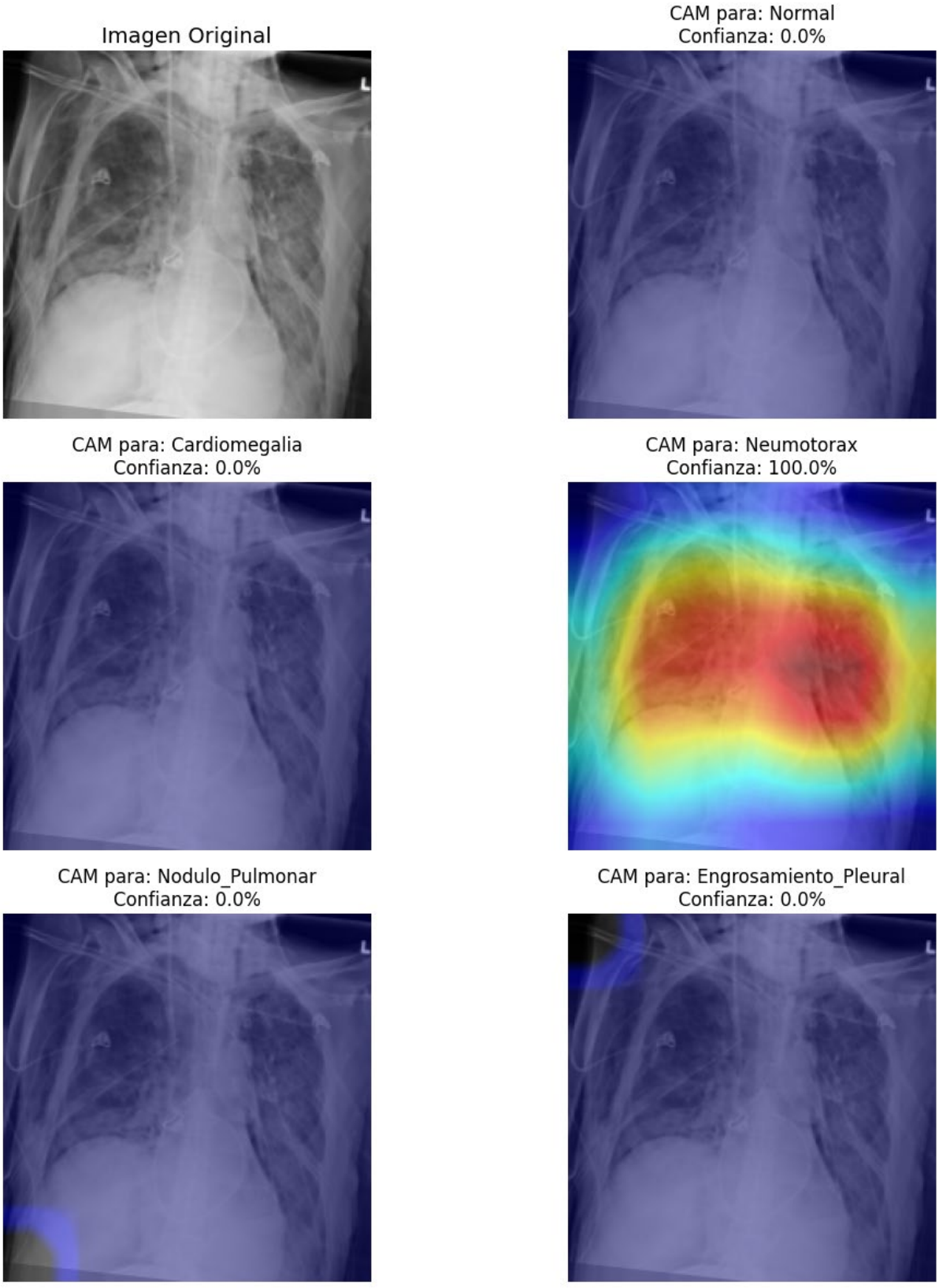
Anexo C 6. *DenseNet121 resultado Grad-CAM - Clase Normal*



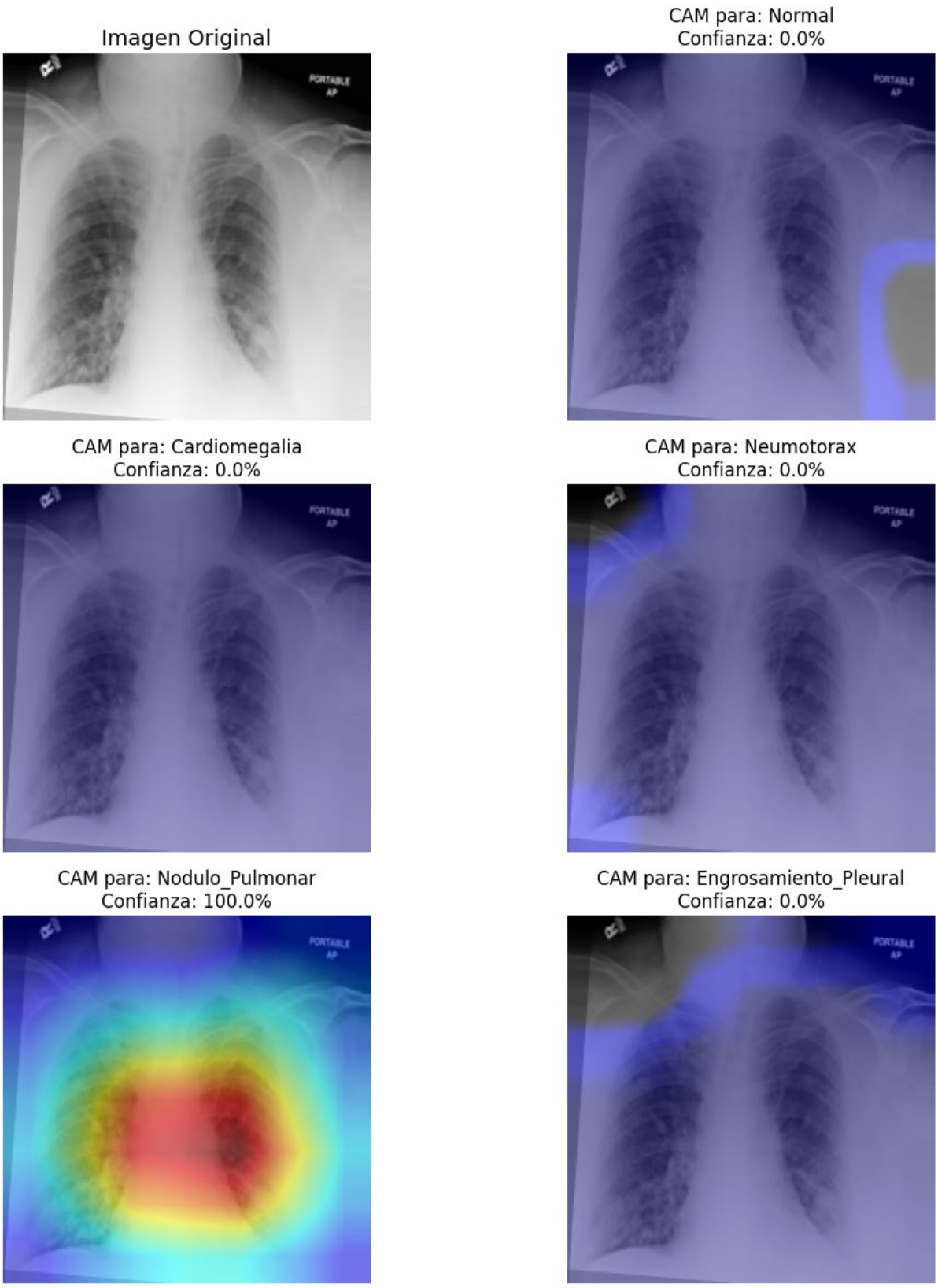
Anexo C 7. *DenseNet121 resultado Grad-CAM - Clase Cardiomegalia*



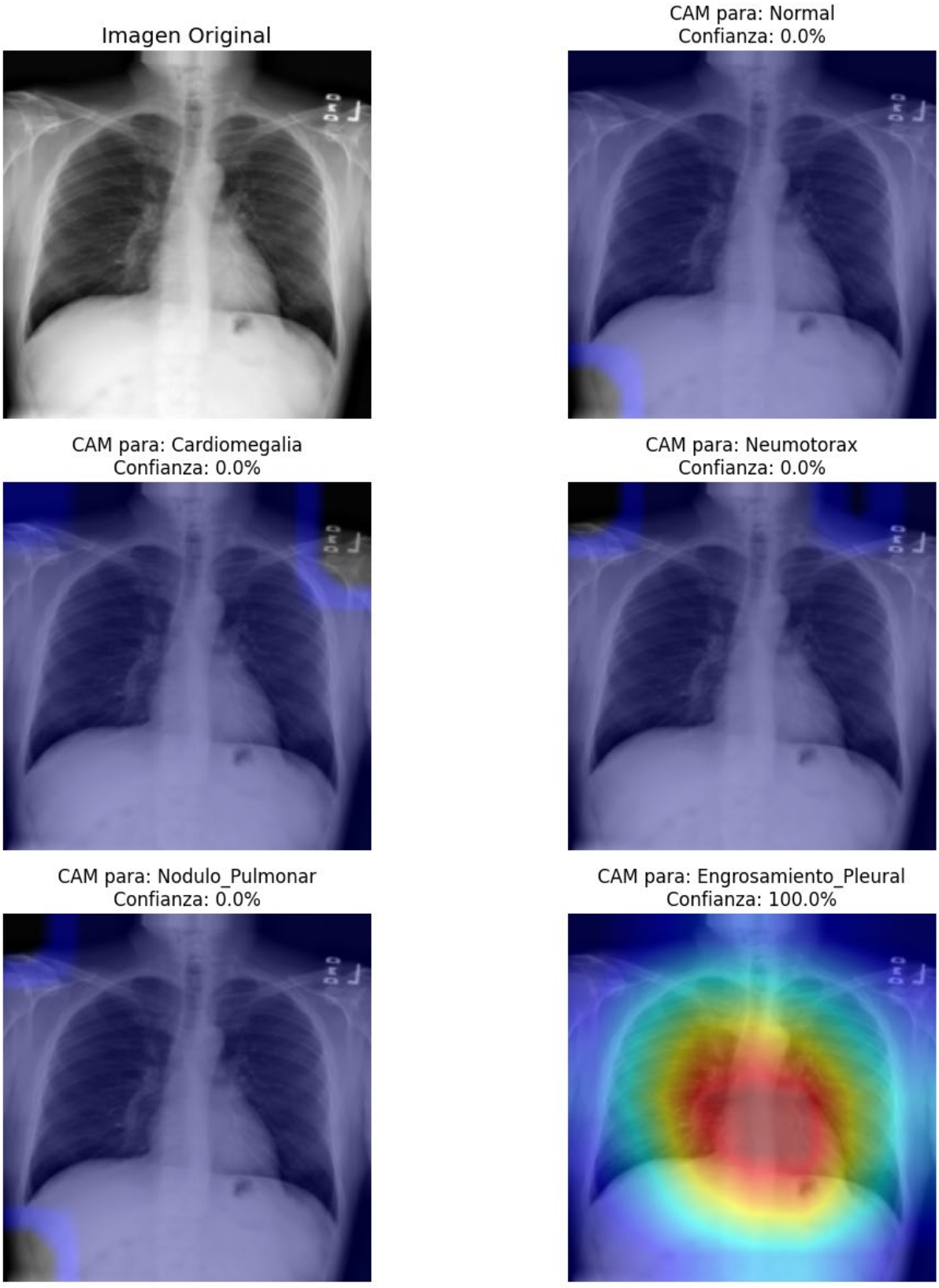
Anexo C 8. *DenseNet121 resultado Grad-CAM - Clase Neumotórax*



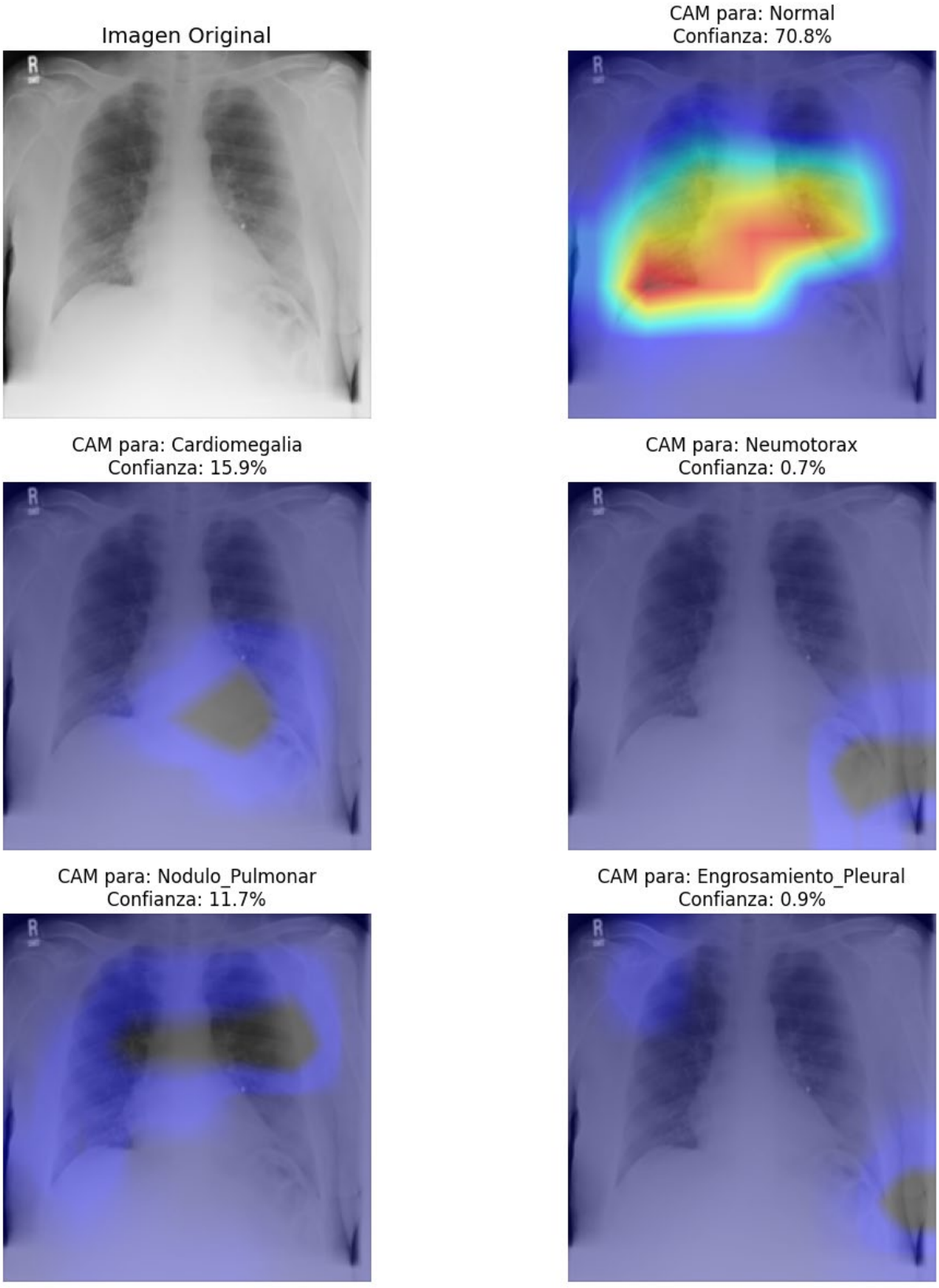
Anexo C 9. *DenseNet121 resultado Grad-CAM - Clase Nódulo Pulmonar*



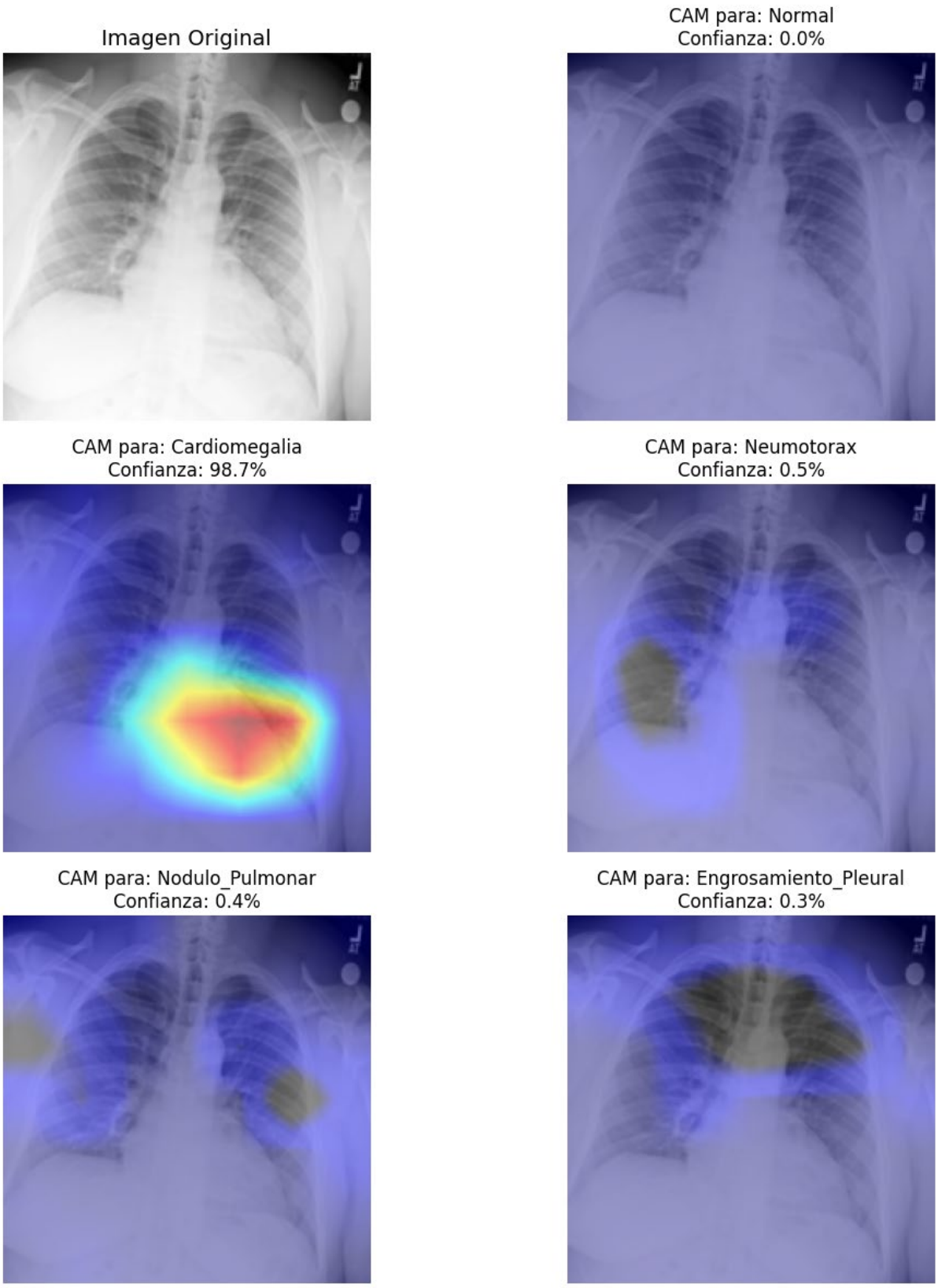
Anexo C 10. *DenseNet121 resultado Grad-CAM - Clase Engrosamiento Pleural*



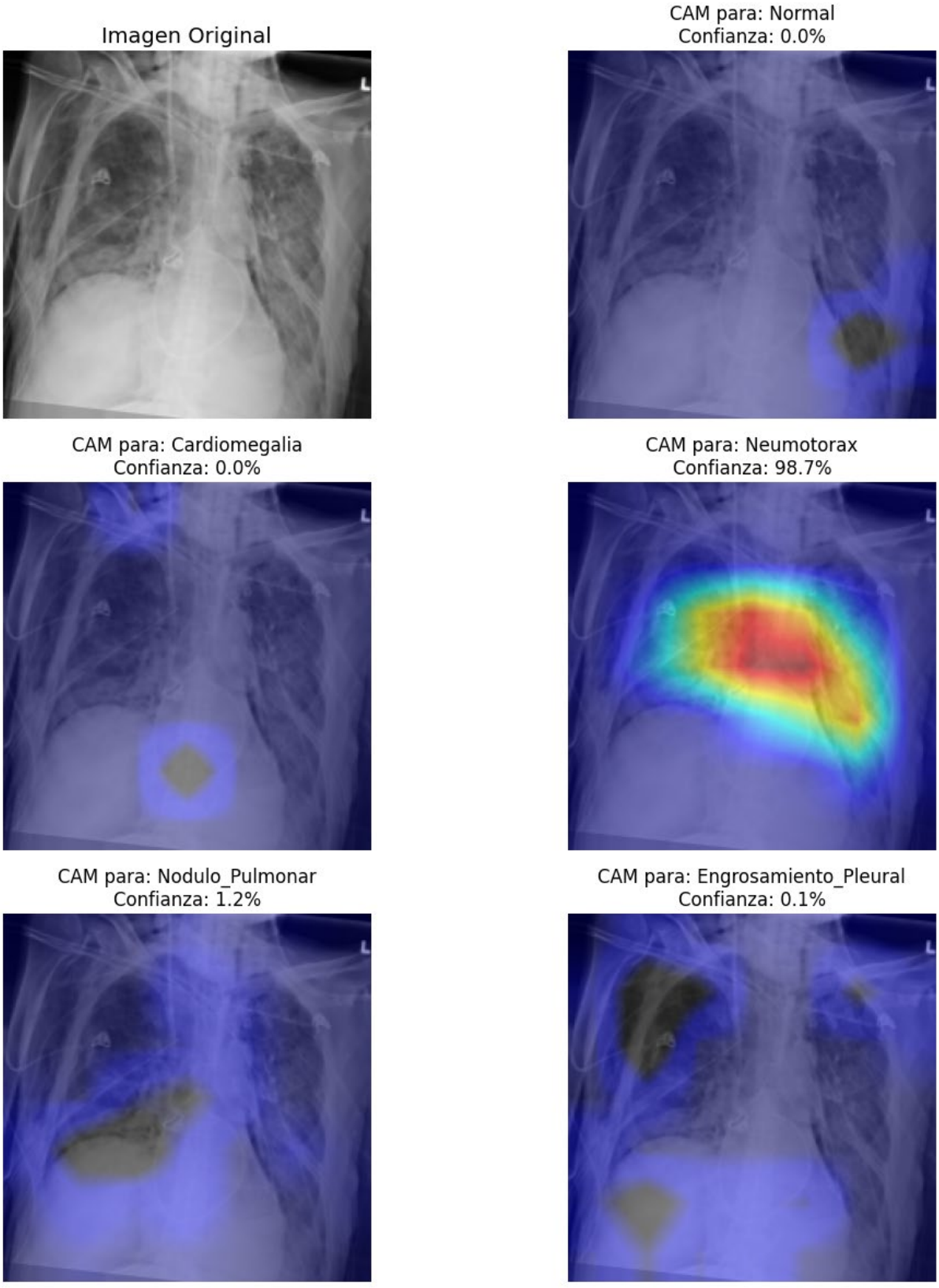
Anexo C 11. *MobileNetV2 resultado Grad-CAM - Clase Normal*



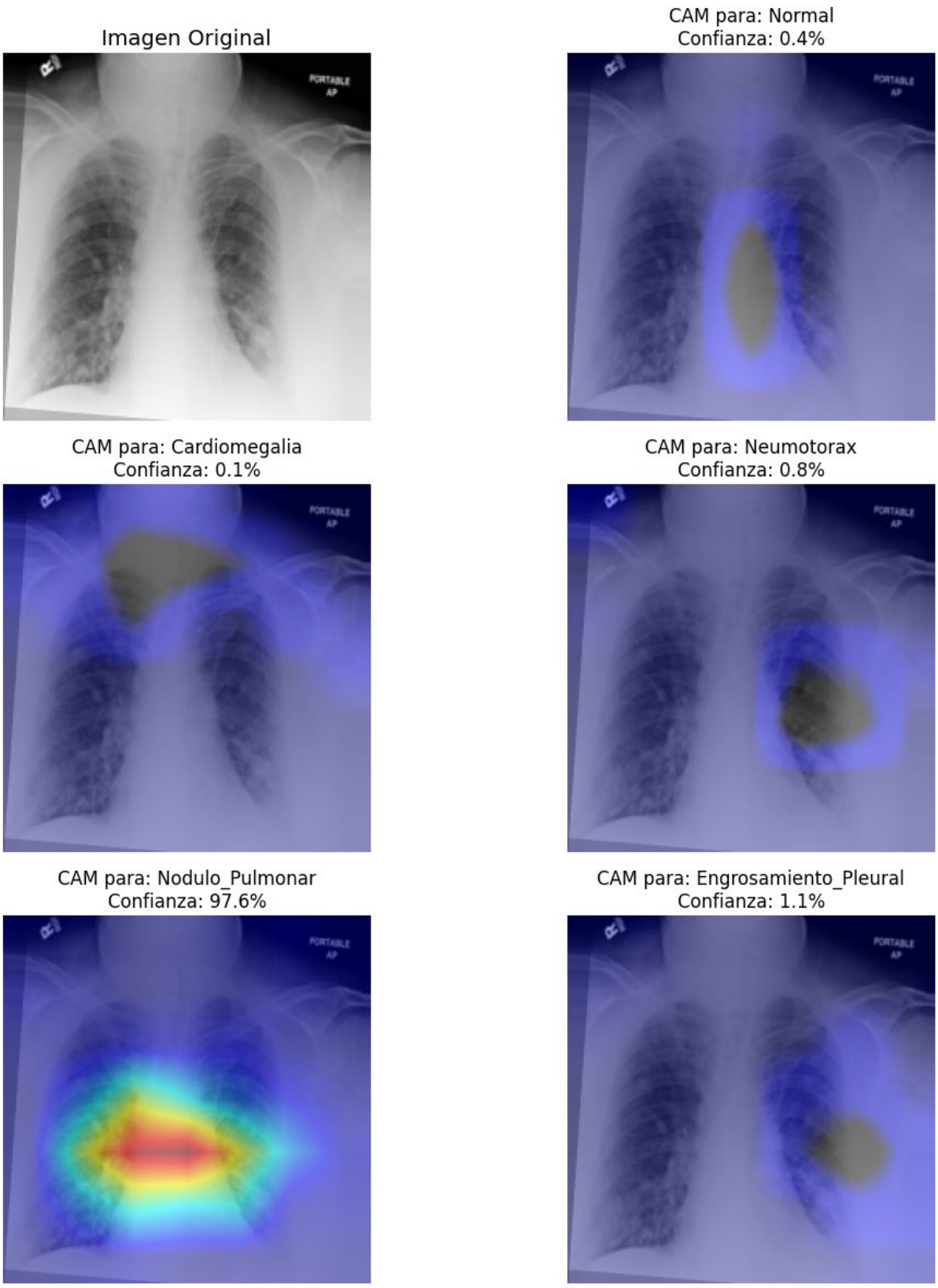
Anexo C 12. *MobileNetV2 resultado Grad-CAM - Clase Cardiomegalia*



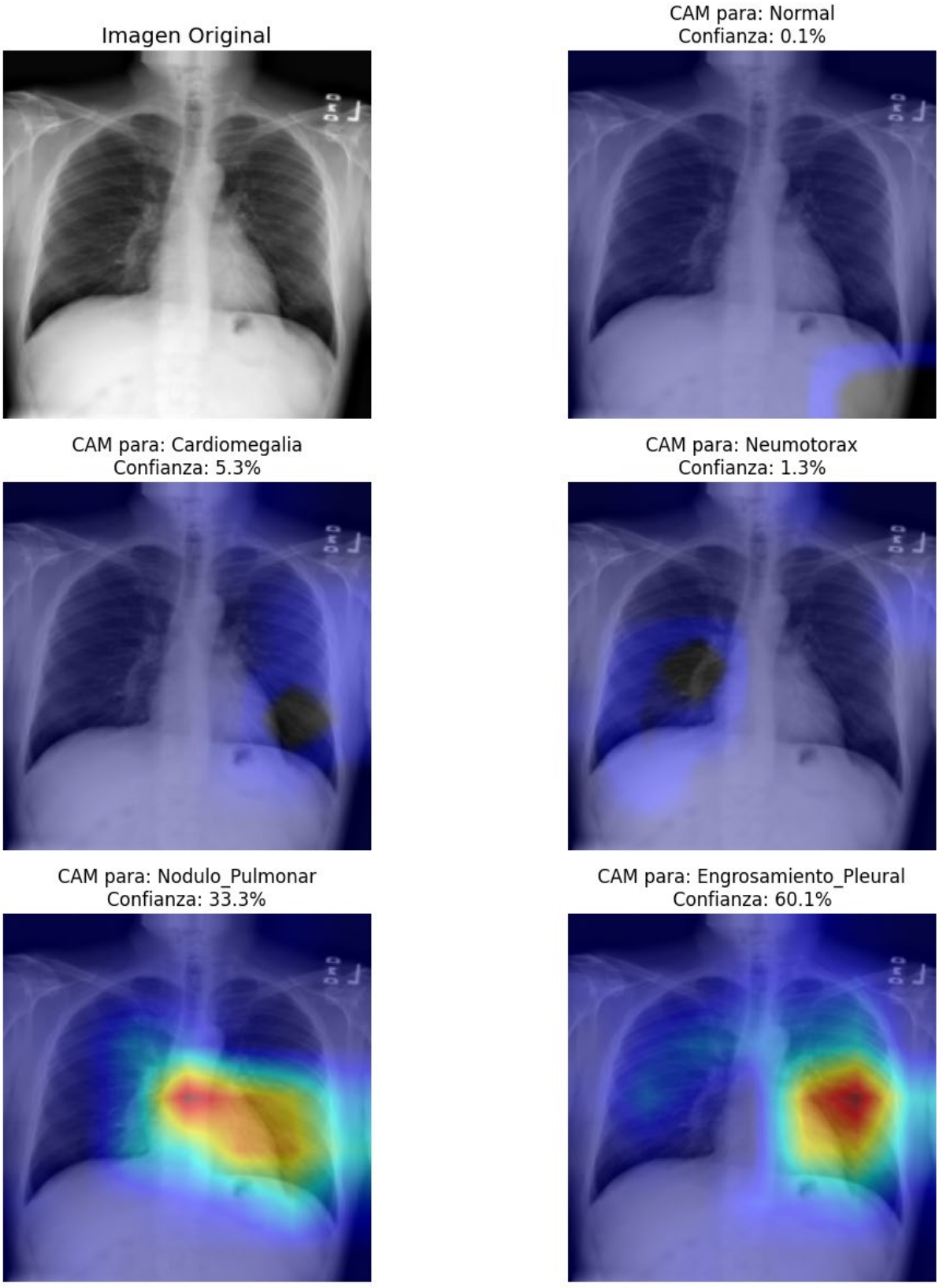
Anexo C 13. *MobileNetV2 resultado Grad-CAM - Clase Neumotórax*



Anexo C 14. *MobileNetV2 resultado Grad-CAM - Clase Nódulo Pulmonar*

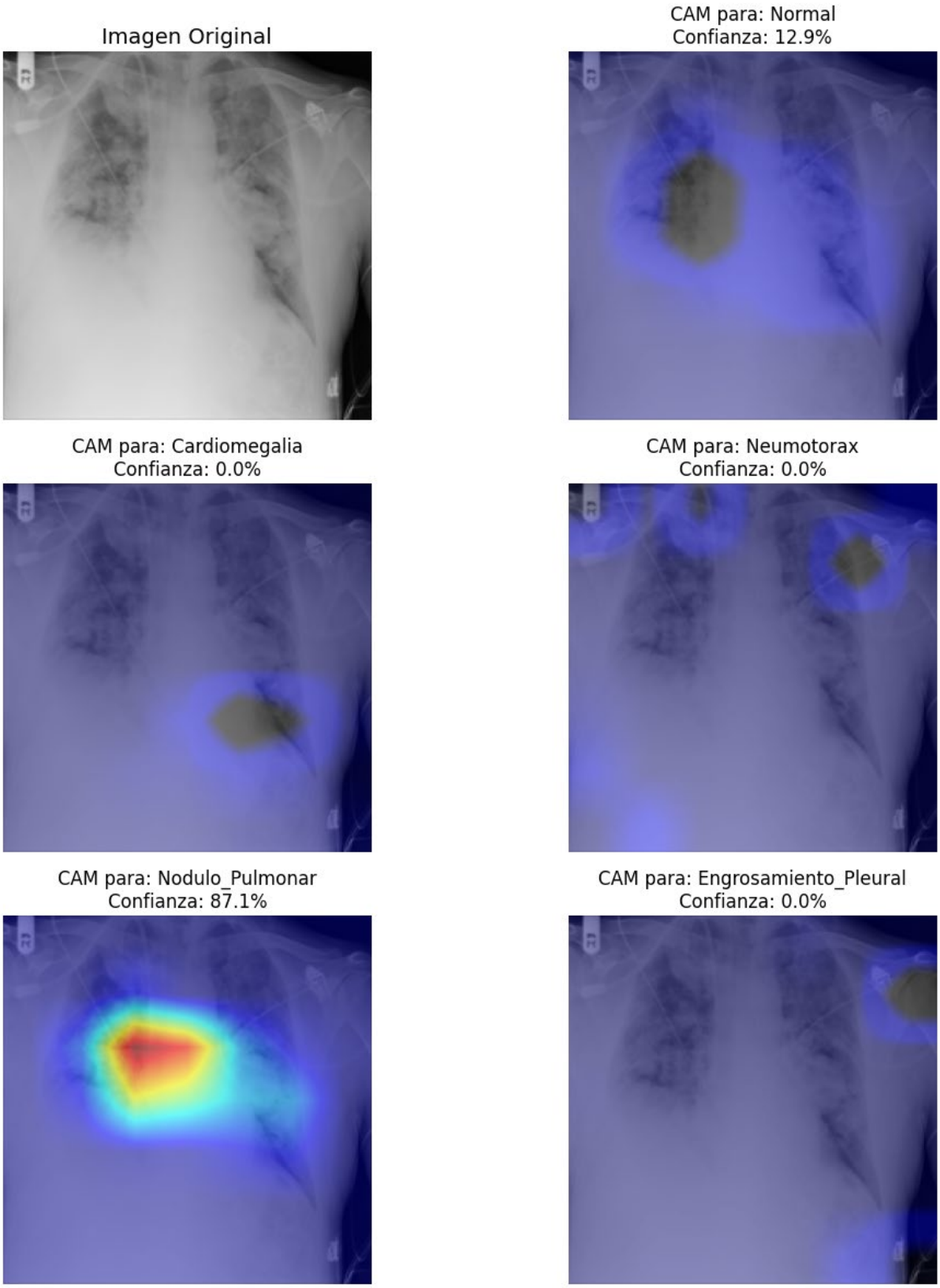


Anexo C 15. *MobileNetV2 resultado Grad-CAM - Clase Engrosamiento Pleural*

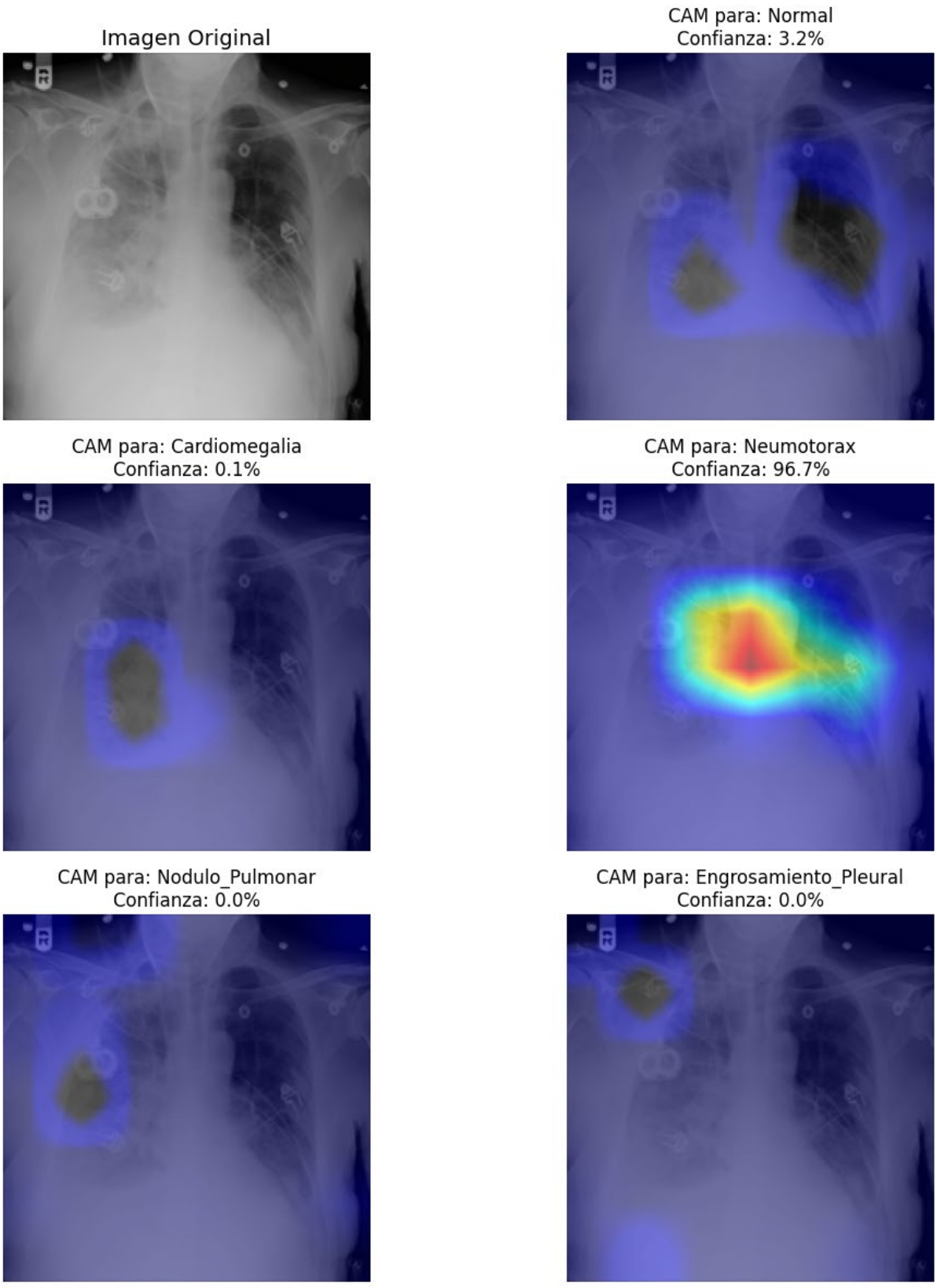


Anexo D. Errores de Predicción Grad-CAM

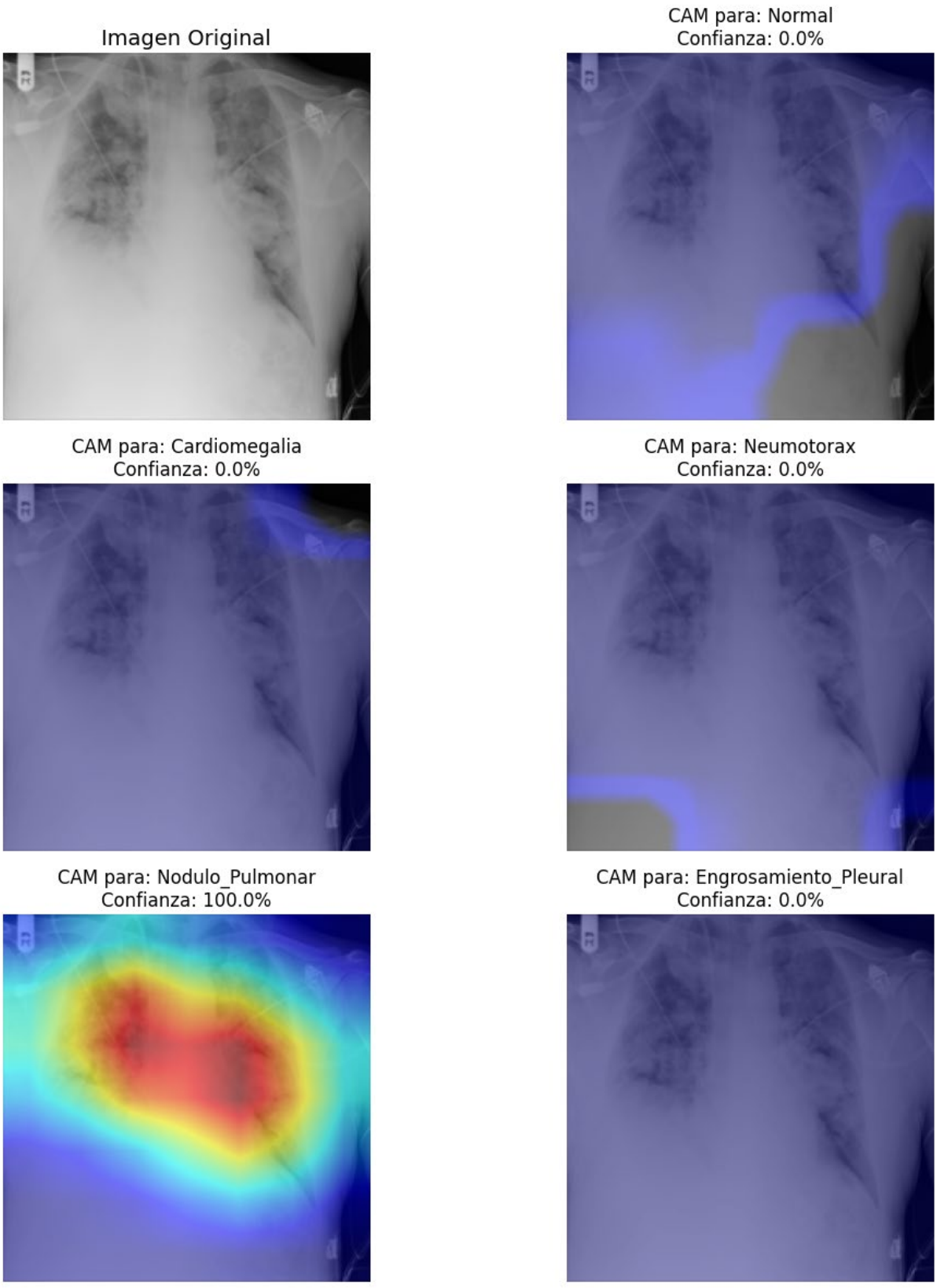
Anexo D 1. *ResNet50 Error de Predicción Grad-CAM - Clase Engrosamiento Normal*



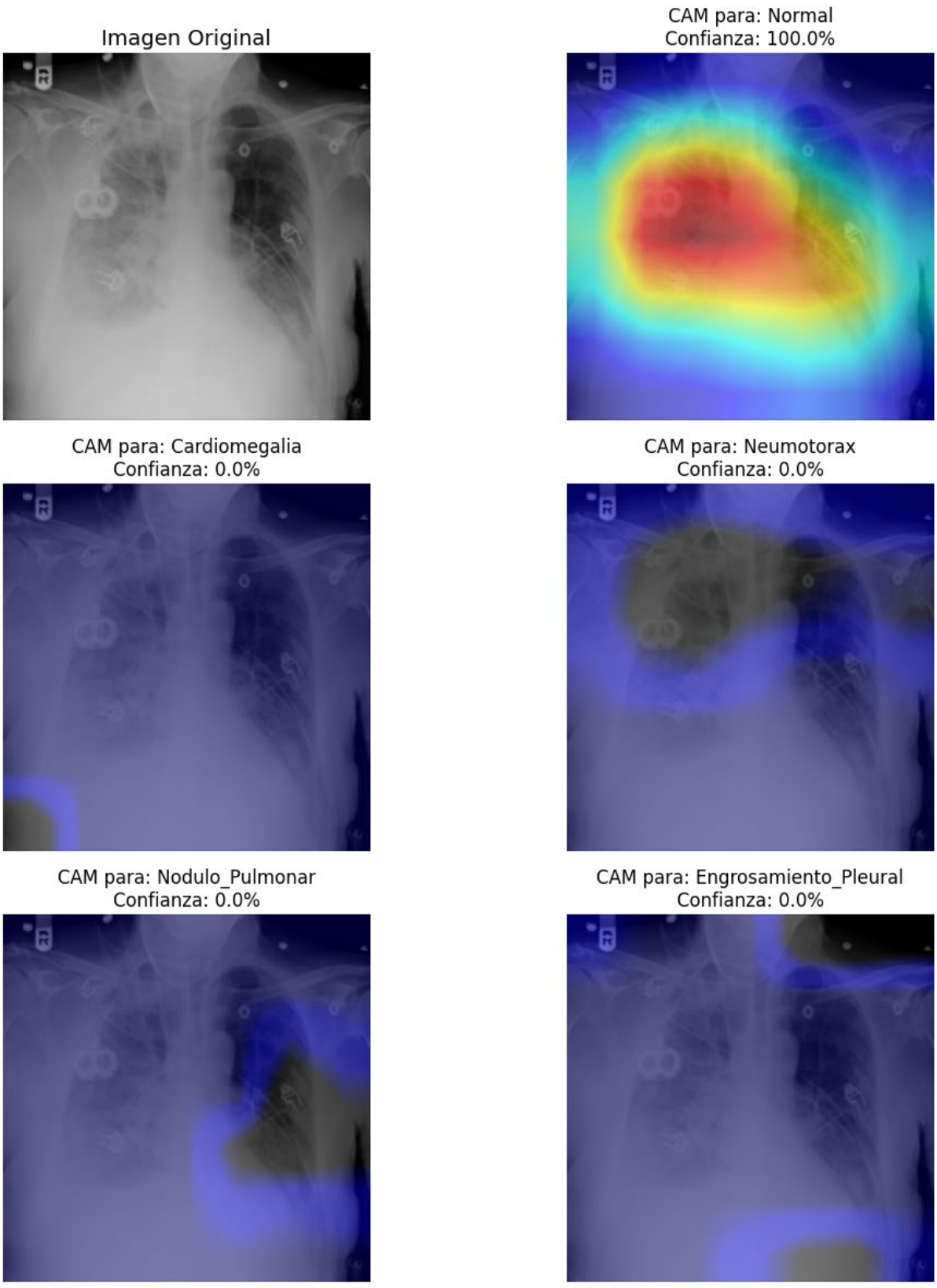
Anexo D 2. *ResNet50 Error de Predicción Grad-CAM - Clase Engrosamiento Neumotórax*



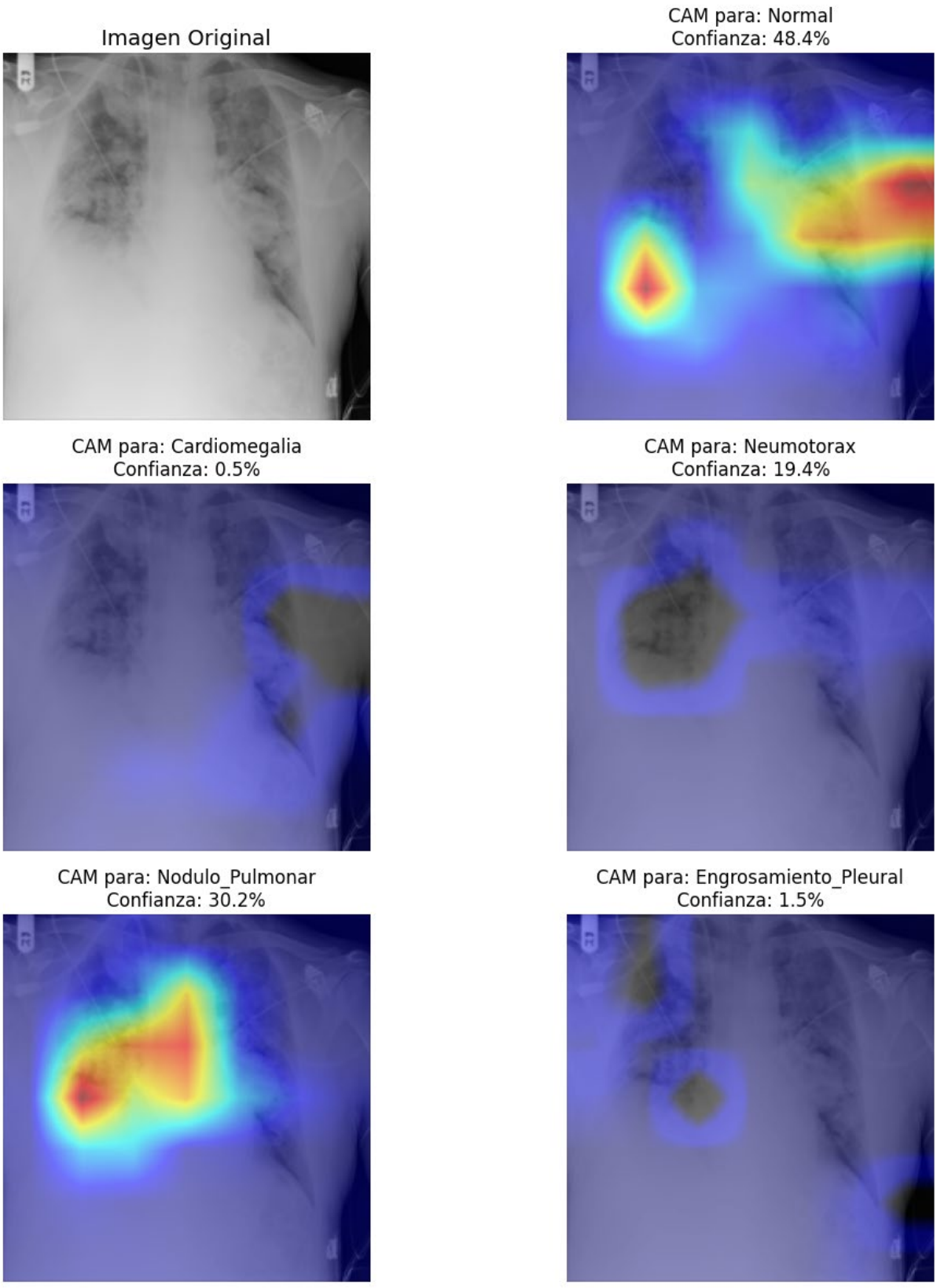
Anexo D 3. *DenseNet121 Error de Predicción Grad-CAM - Clase Engrosamiento Normal*



Anexo D 4. *DenseNet121 Error de Predicción Grad-CAM - Clase Engrosamiento Neumotórax*



Anexo D 5. *MobileNetV2 Error de Predicción Grad-CAM - Clase Engrosamiento Normal*



Anexo D 6. *MobileNetV2 Error de Predicción Grad-CAM - Clase Engrosamiento Neumotórax*