UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

Clasificación Multi-Etiqueta de Patología en Radiografías de Tórax

Miembros:
Rocío Ávalos Morillas y Ainhoa Fraile Pulido

Profesores:
Magda Ruiz, Luis Eduardo Mujica y Santiago Alferez

Asignatura:
Aprendizaje Bioestadístico

Escola d'Enginyeria de Barcelona Est
Universitat Politècnica de Catalunya

15/06/2025

1. OBJETIVO

El objetivo principal de este proyecto es desarrollar y comparar arquitecturas de redes neuronales convolucionales (CNNs) para la clasificación multi-etiqueta de patologías torácicas en radiografías de tórax, utilizando un subconjunto estratégico del dataset NIH ChestX-ray14. Adicionalmente, se explora la estructura interna de las representaciones aprendidas por los modelos mediante técnicas de análisis no supervisado (PCA y K-Means clustering) y supervisado (ROC, Mmatriz de confusión, Hamming).

2. PRINCIPALES HALLAZGOS

2.1. Selección de patologías

Se seleccionaron estratégicamente cuatro patologías que cumplen criterios técnicos y clínicos rigurosos (Figura 1 y 2):

- Infiltration (3,318 casos - 14.73%): patrones de neumonía/infección, hallazgo patológico más frecuente.
- Effusion (2,126 casos - 9.44%): acumulación de líquido pleural, indicador de insuficiencia cardíaca.
- Atelectasis (2,100 casos - 9.32%): colapso pulmonar, complicación postoperatoria común.
- Pneumothorax (881 casos - 3.91%): aire en espacio pleural, condición de emergencia.

Esta selección presenta un desequilibrio moderado (ratio 3.8:1) ideal para el entrenamiento, con un total de 8,425 instancias patológicas distribuidas en 24,999 imágenes del subconjunto analizado.

2.2. Modelo de clasificación multi-etiqueta

2.2.1. Modelo Simple CNN

Red neuronal convolucional implementada completamente desde cero usando PyTorch puro, con arquitectura jerárquica inspirada en el córtex visual humano. Incluye 3 bloques convolucionales progresivos (3→16→32→64 canales), batch normalization, dropout y 425,860 parámetros entrenables. Diseñada para aprendizaje progresivo: detección de bordes → formas → patrones patológicos complejos.

2.2.2. EfficientNet

Arquitectura de Google Research pre-entrenada en ImageNet con 4,012,672 parámetros totales. Utiliza escalado conjunto optimizado de profundidad, ancho y resolución. Implementa transfer learning con extractor convolucional (1,280 características) y clasificador personalizado con dropout del 40%. Entrenamiento con precisión mixta, BCEWithLogitsLoss, optimizador AdamW ($lr=3\times10^{-5}$) y scheduler ReduceLROnPlateau para convergencia estable.

2.2.3. DenseNet

Arquitectura de Facebook AI Research con conexiones densas que preservan características médicas de grano fino. Modelo de referencia en clasificación de radiografías de tórax (usado en CheXNet). Cuenta con 1,024 características antes del clasificador y dropout del 50%. Configuración ultra-conservadora con AdamW ($lr=1\times10^{-5}$, weight decay$=2\times10^{-3}$), scheduler agresivo (reducción 50% cada 2 epoch) y early stopping. Las conexiones densas favorecen la reutilización de características y captura de detalles clínicamente significativos.

3. RENDIMIENTO DE LOS MODELOS

3.1. Parámetros de rendimiento

Para evaluar los modelos se emplearon varias métricas: Hamming Score (precisión promedio por etiqueta), Exact Match Accuracy (predicción perfecta de todas las etiquetas), F1-score promedio y AUC promedio (Figura 3).

- DenseNet-121 presentó el mejor rendimiento general con un Hamming Score de 81.9% y una Exact Match Accuracy de 38.3%, lo que indica alta precisión y capacidad para predecir múltiples etiquetas simultáneamente.
- EfficientNet-B0 se destacó en métricas de balance y discriminación, con el mejor F1-score (0.259) y AUC promedio (0.700), reflejando una mayor capacidad de generalización.
- SimpleCNN, aunque más ligero, obtuvo resultados inferiores en todas las métricas.

3.2. Evaluación no supervisada

El análisis de clustering aplicado a las representaciones latentes de DenseNet-121 reveló hallazgos significativos sobre la estructura interna del conocimiento médico aprendido:

- Captura de varianza y preprocesamiento: el uso de RobustScaler permitió capturar el 99.9% de la varianza total, en contraste con el 59.5% alcanzado por StandardScaler, lo que evidencia la presencia de valores atípicos característicos de imágenes radiológicas. Además, el primer componente principal (PC1) explicó el 99.7% de la varianza, indicando un eje dominante de variabilidad informativa en el espacio latente.
- Estructura geométrica del clustering: el análisis de k-means con Silhouette Score de 0.856 confirmó una estructura de clusters bien definida, con una solución óptima de k=3. Los grupos resultantes mostraron una partición geométrica coherente y un balance adecuado en sus tamaños, reflejando cohesión intra-cluster y separación inter-cluster sobresalientes.
- Correspondencia clínica e interpretación: a pesar del bajo Adjusted Rand Index (ARI = 0.002) respecto a las etiquetas diagnósticas tradicionales, esta desalineación sugiere que el modelo ha aprendido representaciones que capturan relaciones patológicas más complejas, posiblemente no reflejadas en categorías clínicas explícitas. La patología "Infiltration" apareció de forma dominante en los tres clusters (pureza entre 36.8% y 51.0%), lo que refuerza su papel como condición común y co-ocurrente.

Estos hallazgos sugieren que DenseNet-121 desarrolla una representación latente estructurada y clínicamente relevante, con potencial para identificar patrones radiológicos emergentes o subtipos patológicos aún no formalizados en las etiquetas estándar.

3.3. Evaluación Dataset test (pacientes held-out)

La validación final se llevó a cabo utilizando un conjunto de test compuesto por 2,469 radiografías torácicas de 659 pacientes no vistos durante el entrenamiento, lo que permitió evaluar la capacidad de generalización de los modelos en un escenario clínico real. Los resultados mostraron diferencias notables en el rendimiento por patología.

- En el caso de Effusion, EfficientNet logró el mejor desempeño con un F1-score de 0.388, reflejando una buena precisión clínica.
- Para la clase Infiltration, todos los modelos presentaron una sensibilidad elevada, aunque con precisiones moderadas, lo que sugiere que esta condición es difícil de delimitar con claridad debido a su alta frecuencia y superposición con otras patologías.
- En contraste, Atelectasis fue mejor identificada por DenseNet-121, que alcanzó una mayor especificidad, reduciendo así los falsos positivos.
- Finalmente, Pneumothorax representó el mayor desafío diagnóstico debido a su baja prevalencia (2.8%), lo que limitó significativamente el rendimiento de todos los modelos en esta categoría.

3.4 Análisis detallado por matriz de confusión

El análisis detallado de las matrices de confusión permitió comprender mejor las fortalezas y debilidades de cada modelo.

SimpleCNN mostró una elevada sensibilidad para Infiltration (73.7%), pero con una precisión muy baja (19.3%), indicando una alta tasa de falsos positivos. Su desempeño en Effusion fue más equilibrado, aunque moderado, mientras que en Pneumothorax su rendimiento fue crítico, con apenas un 2.9% de precisión y 1.4% de sensibilidad, lo que evidencia una detección prácticamente inexistente.

Por otro lado, EfficientNet logró un mejor balance entre precisión (21.5%) y sensibilidad (70.5%) para Infiltration, y fue el modelo más competente en la detección de Effusion, alcanzando 32.4% de precisión y 48.5% de recall.

DenseNet-121 también mostró un desempeño sólido, especialmente en Infiltration, con una sensibilidad de 68.4% y una precisión de 20.7%,

superando a SimpleCNN en ambas métricas. En el caso de Effusion, evidenció una convergencia casi ideal entre precisión (28.8%) y sensibilidad (28.6%), lo que sugiere una mayor estabilidad diagnóstica.

4. LIMITACIONES DE LOS MODELOS Y POSIBLES MEJORAS FUTURAS

4.1. Limitaciones

A pesar del rendimiento alcanzado por los modelos analizados, existen varias limitaciones:

- Desequilibrio en la distribución de clases: esto repercute negativamente en la capacidad de detección de los modelos. Esto es especialmente relevante en nuestro contexto, ya que la omisión de enfermedades poco frecuentes puede tener consecuencias críticas.
- Complejidad multi-etiqueta del problema: representa un desafío predecir combinaciones clínicas concurrentes en imágenes radiológicas (se observa en el bajo valor de Exact Match Accuracy). La superposición entre diversas condiciones dificulta una clasificación precisa y completamente correcta.
- Subconjunto de datos: se analizaron 25,000 imágenes del total de 112,120 disponibles. Aunque este tamaño es considerable, esta reducción limita la diversidad de casos clínicos considerados.

4.2. Futuras mejoras

Frente a estas limitaciones, futuras líneas de trabajo podrían ser:

- Estrategias de manejo del desbalance de clases.
- Expansión del entrenamiento al Dataset completo.
- Incorporar metadatos clínicos (como edad, sexo o antecedentes del paciente) para contextualizar mejor las predicciones.
- Desarrollar sistemas multimodales, combinando imágenes radiológicas con datos clínicos estructurados y no estructurados.

5. CONCLUSIONES FINALES Y POTENCIALES IMPLICACIONES

Los resultados obtenidos consolidan a DenseNet-121 como la arquitectura más robusta del estudio, alcanzando un Hamming Score del 81.9%. Su comportamiento conservador frente a patologías críticas como el neumotórax (minimizando falsos positivos) refuerza su idoneidad en contextos de alta exigencia clínica.

Por otro lado, EfficientNet-B0 destacó por su excelente equilibrio entre precisión y sensibilidad, con un F1-score promedio de 0.259 y un AUC de 0.700. Esta configuración lo hace particularmente adecuado para tareas de cribado o screening.

Una posible opción es la incorporación de estos modelos en sistemas de triaje automatizado. Esto puede acelerar la priorización de estudios con hallazgos sospechosos, mejorando la eficiencia en servicios de urgencias. Aun así, es importante destacar que debe tener un uso asistencial y no sustitutivo, donde el modelo actúa como herramienta de apoyo. También pueden ofrecen valor en contextos formativos, permitiendo a estudiantes explorar patrones diagnósticos con soporte visual para la interpretación de radiografías.

6. IA generativa

Se utilizaron herramientas de IA generativa (ChatGPT, Claude) para asistencia en:

- Optimización de código: depuración de implementaciones de métricas multi-etiqueta.
- Revisión de literatura: síntesis de trabajos previos en clasificación médica.
- Redacción técnica: mejora de claridad en explicaciones metodológicas.

Ejemplo de prompt utilizado: *"Explica las diferencias entre exact match accuracy y hamming score en clasificación multi-etiqueta médica, con ejemplos prácticos"*

La IA generativa facilitó la comprensión de conceptos técnicos complejos y mejoró la calidad de la documentación científica, sin comprometer la originalidad del análisis experimental.

7. ANEXO

7.1. Documento de análisis y código

El documento IPython Notebook (.ipynb) que contiene el código de análisis, gráficos y resultados se encuentra en el siguiente enlace de GitHub:

[Documento ipynb del segundo trabajo]

7.2. Figuras y tablas

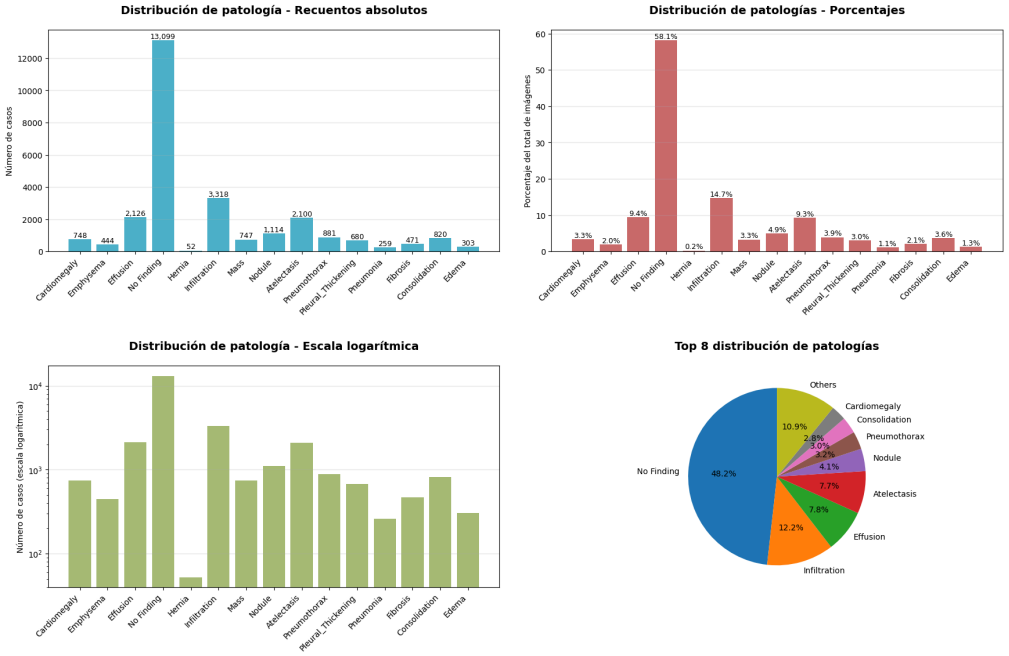


Figura 1. Distribución de las patologías

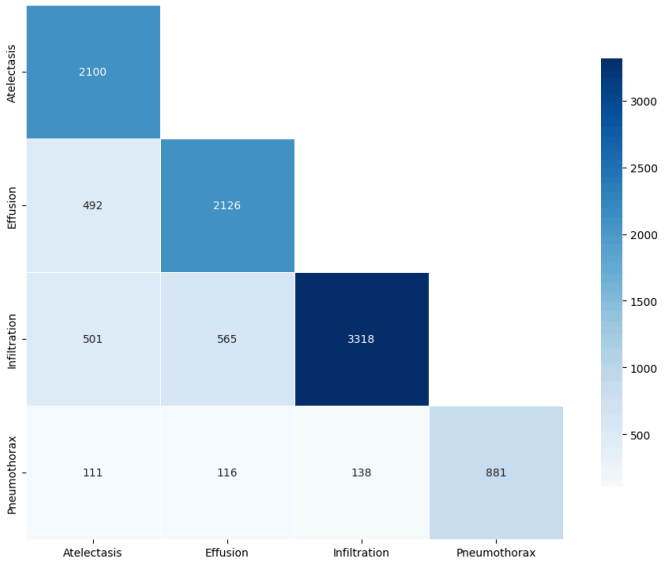
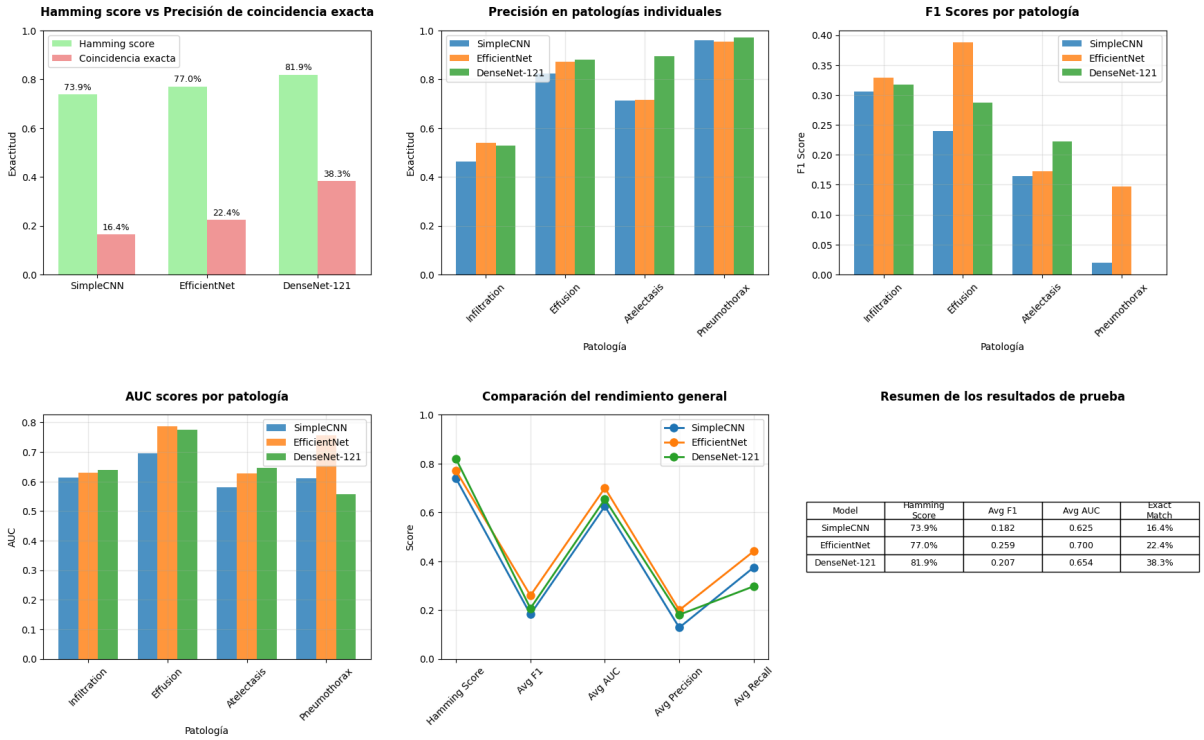


Figura 2. Matriz de co-ocurrencia de las patologías seleccionadas



Model	Hamming Score	Avg F1	Avg AUC	Exact Match
SimpleCNN	73.9%	0.182	0.625	16.4%
EfficientNet	77.0%	0.259	0.700	22.4%
DenseNet-121	81.9%	0.207	0.654	38.3%

Figura 3. Resultados de la evaluación final del conjunto de prueba

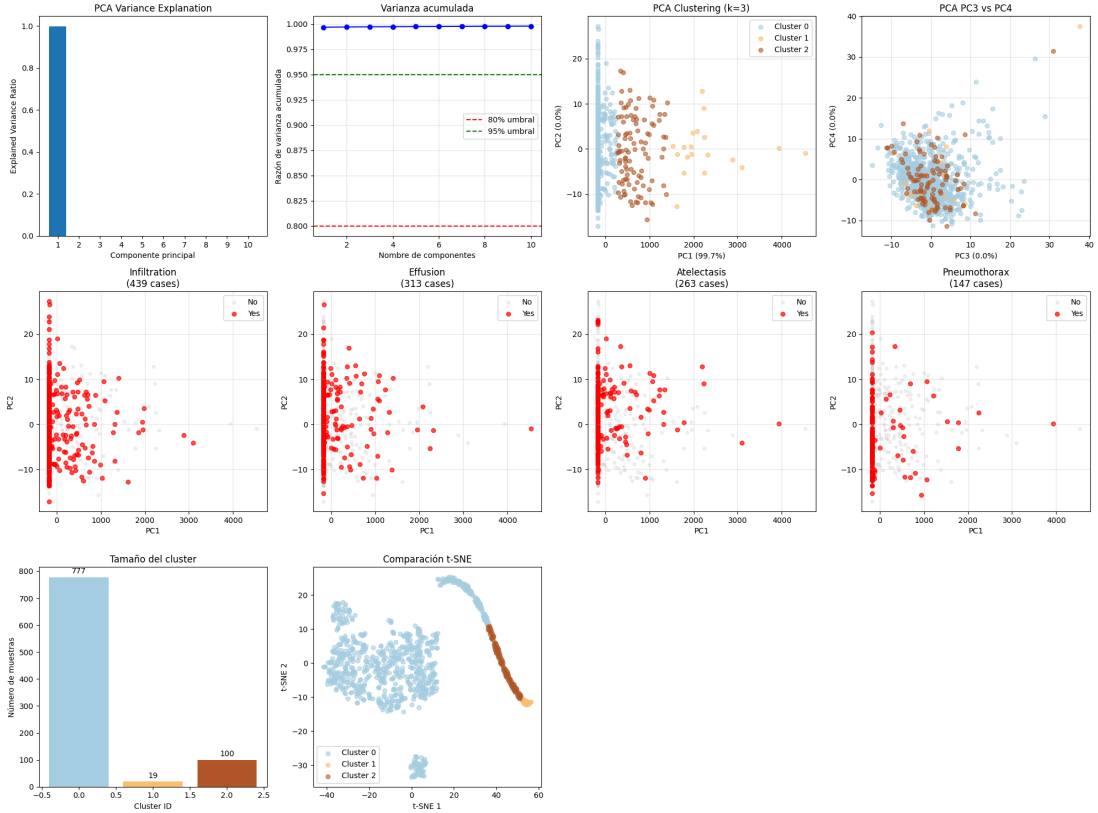


Figura 4. Análisis PCA + K-means completo para funciones de DenseNet

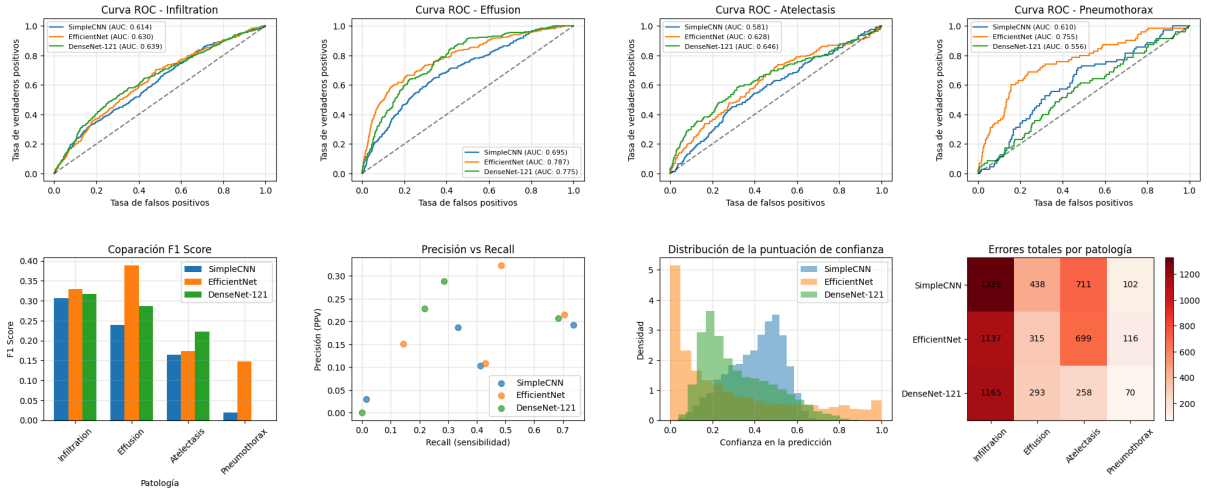


*Figura 5. Análisis de las métricas **ROC** y **F1-score** para las diferentes patologías, utilizando el conjunto **Held-out** y comparando los tres modelos: **SimpleCNN**, **EfficientNet** y **DenseNet**.*