



Detección de neumonía a partir de radiografías de tórax mediante validación cruzada estratificada y Fine-Tuning de DenseNet121

Pneumonia detection from chest X-rays using stratified cross-validation and DenseNet121 Fine-Tuning

Marco Fidel Mayta Quispe^{1,2} , Leonid Alemán Gonzales^{1,2} 

¹ Universidad Nacional del Altiplano. Facultad de Ingeniería Estadística e Informática. Puno, Perú.

² Universidad Nacional del Altiplano. Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos (IIICCD). Puno, Perú.

DOI: 10.64966/ingeniare.v34.02

Resumen

La neumonía es una enfermedad muy común, en especial en zonas altoandinas, su diagnóstico muchas veces suele ser lento por la falta de tecnología o personal, por ello es primordial desarrollar un modelo de Inteligencia Artificial que pueda diagnosticar neumonía mediante placas radiográficas, en este trabajo se entrenó un modelo optimizado para el diagnóstico de neumonía a partir de radiografías de tórax, utilizando Redes Neuronales Convolucionales (CNN) junto con validación cruzada estratificada. La base de datos que se utilizó fue *Chest X-ray Images* (Pneumonia) de acceso público en Kaggle, el cual fue preprocesado y balanceado para disminuir sesgos de clase. Se seleccionó la arquitectura *DenseNet-121*, la cual fue ajustada mediante fine tuning, regularización con *dropout*, y normalización por lotes. Posteriormente se evaluó mediante validación cruzada estratificada con cinco particiones. El modelo mostró un rendimiento promedio excelente, con un valor medio de AUC-ROC de 0,9967 y F1 elevados, especialmente para la clase neumonía, lo que evidencia una alta capacidad de discriminación entre casos normales y patológicos. Estos resultados superan otros resultados por enfoques similares sin validación estratificada ni ajuste fino. Se concluye que la combinación de arquitecturas preentrenadas, técnicas modernas de regularización y una evaluación estadística rigurosa permite mejorar significativamente el diagnóstico automático de neumonía, especialmente útil en regiones con acceso limitado a especialistas o tecnología avanzada.

Palabras clave: Diagnóstico automático, neumonía, imágenes médicas, DenseNet, AUC-ROC, validación cruzada estratificada.

Abstract

Pneumonia is a very common disease, especially in high Andean areas, its diagnosis is often slow due to the lack of technology or personnel, therefore it is essential to develop an Artificial Intelligence model that can diagnose pneumonia using X-ray plates, in this work, an optimized model was trained for the diagnosis of pneumonia from chest X-rays, using convolutional neural networks (CNN) together with stratified cross-validation. The database used was Chest X-ray Images (Pneumonia), publicly accessible on Kaggle, which was preprocessed and balanced to reduce class biases. The *DenseNet-121* architecture was selected and adjusted using fine-tuning, dropout regularization, and batch normalization. It was subsequently evaluated using stratified cross-validation with five partitions. The model showed excellent performance, with an average AUC-ROC of 0,9967 and high F1-scores, especially for the pneumonia class, demonstrating strong ability to discriminate between normal and pathological cases. These results outperform other results from similar approaches without stratified validation or fine-tuning. It is concluded that the combination of pre-trained architectures, modern regularization techniques, and rigorous statistical evaluation significantly improves the automatic diagnosis of pneumonia, particularly in regions with limited access to specialists or advanced technology.

Keywords: Automated diagnosis, pneumonia, medical imaging, DenseNet, AUC-ROC, stratified cross-validation.

INTRODUCCIÓN

La neumonía es una enfermedad respiratoria aguda causada por diversos agentes infecciosos, incluyendo bacterias, virus y hongos, que provoca inflamación en los alvéolos pulmonares. A nivel global, representa una de las principales causas de morbilidad y mortalidad, especialmente en niños menores de cinco años, adultos mayores y personas con comorbilidades o inmunodepresión. Según la Organización Mundial de la Salud (OMS), la neumonía es responsable de aproximadamente el 15% de todas las muertes de niños menores de cinco años, con cifras que superan los 800.000 fallecimientos anuales. En regiones como América Latina, África y el sudeste asiático, el impacto es aún mayor debido a la limitada disponibilidad de equipos de diagnóstico avanzado y a la escasez de personal médico especializado. El diagnóstico clínico de neumonía suele basarse en la evaluación de síntomas, examen físico, análisis de laboratorio y estudios por imágenes. Las Radiografías de Tórax (RX) continúan siendo la herramienta más accesible y utilizada para confirmar la presencia de infiltrados pulmonares. Sin embargo, la interpretación de estas imágenes requiere experiencia significativa y puede verse afectada por la subjetividad del radiólogo, la calidad de la imagen y la presencia de otras condiciones pulmonares que simulan los hallazgos radiológicos de la neumonía.

En los últimos años, el desarrollo de técnicas de Inteligencia Artificial (IA), en particular el aprendizaje profundo (*Deep Learning*), ha revolucionado el análisis automático de imágenes médicas [8]. Las Redes Neuronales Convolucionales (CNNs, por sus siglas en inglés) han demostrado una capacidad notable para extraer características jerárquicas de imágenes radiográficas [1], alcanzando resultados comparables e incluso superiores a los de expertos humanos en tareas específicas de diagnóstico. Recientes evidencias muestran que los modelos ViT puros ya pueden alcanzar un rendimiento de nivel alto en detección de neumonía [9], variantes ligeras de ViT reducen aún más el coste computacional al tiempo que preservan el modelado de dependencia a largo alcance [10], se han propuesto métodos de adaptación de dominios entre poblaciones para mitigar los efectos de cambio de conjunto de datos [11], varios estudios han abordado la detección automatizada de neumonía mediante CNNs utilizando conjuntos de datos públicos como el “Chest X-ray Images (Pneumonia)” de Kaggle, sin embargo, muchos de estos trabajos no abordan adecuadamente los desafíos metodológicos inherentes, tales como el desbalance de clases, el sobreajuste (*overfitting*) o la falta de validación robusta.

Por lo expuesto, el presente estudio propone un modelo optimizado para el diagnóstico automático de neumonía mediante imágenes de rayos X, empleando una arquitectura convolucional preentrenada de última generación, *DenseNet121*, con técnicas de *fine-tuning*, aumento de datos, balanceo de clases y una evaluación rigurosa mediante validación cruzada estratificada. De esta manera no solo se busca mejorar la precisión de diagnóstico, sino que también obtener una evaluación más confiable y general del desempeño del modelo en diferentes subconjuntos de la población.

Así mismo, es de vital importancia desarrollar trabajos de este tipo y convertirlos en herramientas aplicables en entornos con recursos limitados, donde la disponibilidad de especialistas en radiología es insuficiente para atender la demanda clínica, o no se cuenta con infraestructura tecnológica avanzada, como tomógrafos computarizados. De este modo, la presente investigación no solo busca aportar a la comunidad científica en el campo del diagnóstico asistido por IA, sino también contribuir al desarrollo de soluciones prácticas, accesibles y eficaces en la lucha global contra enfermedades respiratorias como la neumonía.

ANTECEDENTES

La neumonía es una infección respiratoria aguda y una de las principales causas de morbilidad y mortalidad a nivel mundial, especialmente en niños menores de cinco años y en adultos mayores [12], [13]. Según la Organización Mundial de la Salud, en 2018 se registraron alrededor de 1,4 millones de defunciones por neumonía en menores de cinco años [12]. El diagnóstico suele apoyarse en estudios de imagen, especialmente en la radiografía de tórax, que continúa siendo la técnica de primera elección por su disponibilidad y bajo costo [12], [14]. En estas radiografías se observan opacidades pulmonares o consolidaciones típicas de neumonía, sin embargo, en la práctica los hallazgos pueden ser discretos o confundirse con otras entidades, lo que introduce variabilidad en la interpretación, incluso entre radiólogos con experiencia [12], [15], [13]. En servicios de urgencias, donde muchas veces no hay radiólogos de forma permanente, se han documentado discrepancias importantes entre lectores al momento de identificar neumonía en una placa radiográfica [16], [17]. Estos hechos han impulsado el desarrollo de herramientas computacionales que sirvan como apoyo al diagnóstico, con la expectativa de mejorar la eficiencia y eficacia en la detección de neumonía.

En los últimos años, el aprendizaje profundo o *Deep learning* se ha consolidado como una de las principales aproximaciones para analizar radiografías de tórax con fines diagnósticos [12], [1]. A diferencia de los métodos clásicos de visión por computador, que dependen de una selección manual de características, las redes neuronales convolucionales aprenden directamente de los píxeles de la imagen y son capaces de capturar patrones complejos vinculados con la presencia de neumonía [11], [13]. Varios trabajos han mostrado que estos modelos alcanzan niveles altos de exactitud y ayudan a reducir la subjetividad inherente a la lectura humana [12], [13]. El autor [13] utilizó arquitecturas relativamente sencillas de CNN combinadas con aumento de datos y obtuvieron exactitudes cercanas al 90.7% y 93% en la clasificación de radiografías pediátricas normales y con neumonía. El dataset [18], conocido como *pediatric chest X-ray dataset* o *Kermany dataset* que contiene aproximadamente 5200 imágenes de entrenamiento y 624 de prueba etiquetadas como neumonía y sano, se ha convertido en un punto de referencia recurrente para evaluar modelos en poblaciones pediátricas.

Un factor importante para el avance de estos métodos ha sido la disponibilidad de grandes bases de datos etiquetadas. En 2017, [19] presentaron *ChestX-ray8* posteriormente ampliado a *ChestX-ray14*, el cual reúne más de 100.000 radiografías frontales con 14 diagnósticos posibles, incluyendo neumonía. En base a esos datos, [1] entrenaron CheXNet, una CNN basada en DenseNet-121 que mostró un rendimiento comparable al de radiólogos en la clasificación de neumonía frente a no neumonía y un AUC-ROC cercano a 0,90 en validación interna, superando

los resultados sobre desempeño promedio de cuatro radiólogos [1], [15]. El algoritmo se entrenó y evaluó tomando como referencia la lectura radiológica convencional, lo que puede llevar a una sobreestimación del rendimiento real si esta lectura no representa un estándar de oro [12]. Posteriormente se presentaron bases de datos aún más grandes y con esquemas de anotación más elaborados, tales como *CheXpert*, el cual contiene 224.316 radiografías anotadas para 14 hallazgos torácicos, entre ellos neumonía, e incorpora distintas categorías de incertidumbre en las etiquetas [20]. MIMIC-CXR es un repositorio con 377.110 radiografías de tórax y con informes radiológicos, pensado específicamente para investigación clínica [14].

Con respecto a las arquitecturas empleadas, los primeros trabajos se apoyaron especialmente en CNN desarrolladas originalmente para tareas generales de clasificación de imágenes, tales como *ResNet* [21] en el diagnóstico de neumonía pediátrica con un número relativamente reducido de imágenes. Así como también, *VGG16/VGG19*, *inceptionv3* y *Xception* se han usado como bases para ajustar modelos específicos en radiografías de tórax [11], [13]. Una de las arquitecturas que más atención ha recibido es *DenseNet-121*, además de su empleo en *CheXNet*, ha sido probado en otros trabajos por su capacidad de reutilizar características entre capas y mitigar el problema del desvanecimiento del gradiente [1], [22]. En años recientes aparecieron familias de redes como *EfficientNet*, que buscan un escalamiento más sistemático en profundidad, anchura y resolución. En la validación interna de [12], el modelo alcanzó un AUC-ROC cercano a 0,98, sin embargo, cuando se evaluó en cohortes externas con diagnósticos confirmados por un panel clínico y tomografía computarizada, el AUC se redujo a aproximadamente 0,74 [12].

Al mismo tiempo han surgido propuestas basadas en Vision Transformers (ViT) para el análisis de imágenes médicas. Estos modelos, inspirados en los transformadores usados en procesamiento de lenguaje natural, dividen la imagen en parches y utilizan mecanismos de autoatención para capturar relaciones de mayor alcance. Mientras las CNN tienden a explotar bien la información local, los ViT pueden modelar dependencias globales en la imagen [9]. En el caso específico de neumonía, [9] aplicaron un modelo tipo Vision Transformer al conjunto pediátrico de Kermany y reportaron una exactitud del 97,6 %, con 95 % de sensibilidad y 98 % de especificidad [9]. En sus experimentos, el transformer superó el rendimiento de varias CNN entrenadas bajo las mismas condiciones, lo que sugiere que este tipo de arquitecturas puede aportar ventajas en la interpretación de radiografías, aunque su uso en entornos clínicos todavía es incipiente

Otra línea de trabajo relevante es el uso de modelos de conjunto (*ensembles*). En lugar de depender de una sola red, se combinan las predicciones de varios modelos para aprovechar sus puntos fuertes. Los autores [11], propusieron un sistema que integra *GoogLeNet*, *ResNet-18* y *DenseNet-121* para clasificar radiografías en “neumonía” y “normal”. El peso de cada modelo en la combinación final se ajustó utilizando métricas de rendimiento como la precisión, la sensibilidad, el *F1-score* y el AUC. En el conjunto pediátrico de Kermany, el ensemble alcanzó una exactitud del 98,8 % y una sensibilidad también del 98,8 %; en el conjunto RSNA Pneumonia Detection Challenge, que incluye alrededor de 30.000 radiografías de adultos, el mismo enfoque consiguió una exactitud de 86,9 % y una sensibilidad cercana al 87 % [11]. La diferencia entre ambos escenarios refleja la mayor complejidad y heterogeneidad de la población adulta y de los hallazgos en ese tipo de imágenes, pero al mismo tiempo muestra que la combinación de modelos puede mejorar el rendimiento en contextos exigentes y servir como base para esquemas de explicabilidad y localización de lesiones, como Grad-CAM.

También se ha insistido en la necesidad de utilizar métricas que reflejen mejor la calidad del modelo que la simple exactitud global, especialmente cuando existe desbalance entre clases. La curva ROC y el AUC son indicadores frecuentes para evaluar el comportamiento del clasificador en distintos umbrales; se complementan habitualmente con la sensibilidad, la especificidad y el F1-score, que resume el compromiso entre falsos positivos y falsos negativos [11], [13]. Además, muchos estudios emplean validación cruzada con varios pliegues 5 o 10, según el caso, para disminuir la dependencia de una sola partición entrenamiento/prueba y obtener estimaciones más estables del rendimiento promedio [11]. Se considera una buena práctica definir de forma nítida los conjuntos de entrenamiento, validación y prueba, idealmente separando pacientes para evitar fugas de información, y acompañar los resultados numéricos con matrices de confusión que ayuden a entender en qué tipos de casos se producen más errores.

La utilidad de estos modelos en la práctica clínica se ha empezado a estudiar mediante comparaciones directas con radiólogos. En trabajos retrospectivos, se ha observado que, cuando solo se dispone de la imagen, las CNN entrenadas con grandes bases de datos alcanzan sensibilidades similares a las de especialistas para detectar neumonía en CXR [12], [15]. En algunos casos, como en el estudio de [15] se ha señalado que el algoritmo podía igualar o superar el rendimiento de los radiólogos en determinadas tareas. Sin embargo, en la práctica diaria, el diagnóstico se apoya también en la clínica, los análisis de laboratorio y el seguimiento, por lo que el papel del modelo no es reemplazar al médico, sino complementar su juicio. En esta línea, [17] mostraron que un modelo de *deep learning* que detecta múltiples hallazgos en CXR mejoró la exactitud de los radiólogos en un estudio multilector y multicaso. Los autores [16], llegaron a conclusiones similares al evaluar una herramienta comercial de apoyo a la lectura de radiografías, que permitió aumentar el rendimiento de los lectores y reducir el tiempo de lectura. Al mismo tiempo, estudios como el de [13] recuerdan que, cuando se emplean estándares diagnósticos más estrictos (como un diagnóstico final consensuado apoyado en tomografía), el rendimiento del modelo puede ser más modesto que el observado en validaciones internas, aunque siga siendo clínicamente útil.

METODOLOGÍA

Diseño del estudio

Este estudio se basó en un enfoque experimental cuantitativo, orientado a la evaluación del rendimiento de un modelo de red neuronal convolucional profunda para la detección automatizada de neumonía en imágenes radiográficas de tórax. El pipeline metodológico incluyó la recolección de datos, preprocesamiento, diseño e implementación del modelo, validación cruzada estratificada, balanceo de clases, aumento de datos, entrenamiento supervisado, y análisis estadístico de resultados.

Conjunto de datos

Se utilizó el conjunto de datos Chest X-ray Images (Pneumonia) disponible libremente en *Kaggle*, el cual contiene 5,863 imágenes de rayos x clasificadas por expertos médicos en dos categorías: NORMAL (sin signos de infección pulmonar) y PNEUMONIA (infección confirmada). Para garantizar la imparcialidad en la validación, solo se empleó el conjunto de entrenamiento (train) del

dataset, lo que permitió realizar una buena evaluación mediante validación cruzada estratificada.

Entorno de desarrollo

El desarrollo del modelo se realizó íntegramente en Python utilizando bibliotecas especializadas en visión por computador y aprendizaje profundo, incluyendo *TensorFlow/Keras*, *OpenCV*, *NumPy*, *Pandas*, *Scikit-learn* y *Matplotlib*. Las imágenes del dataset se cargaron con *OpenCV* en escala de grises y posteriormente se redimensionaron a 224×224 píxeles, normalizando su rango de intensidad a valores entre 0 y 1. Para garantizar la compatibilidad con la arquitectura *DenseNet121*, las imágenes se transformaron a formato RGB por duplicación de canales. Asimismo, se aplicó aumento de datos mediante rotaciones aleatorias, zoom, desplazamientos y volteos horizontales utilizando *ImageDataGenerator*, con el objetivo de mejorar la generalización del modelo.

Dado el desbalance natural del conjunto de datos, se implementó un *oversampling* de la clase minoritaria mediante remuestreo aleatorio para equilibrar la cantidad de imágenes normales y con neumonía en cada pliegue de validación cruzada. El entrenamiento se ejecutó con el optimizador Adam ($\text{lr} = 1\text{e-}4$) y la función de pérdida *binary crossentropy*, incorporando *EarlyStopping* y *ReduceLROnPlateau* para prevenir sobreajuste y garantizar estabilidad durante el ajuste fino. Todo el código utilizado es reproducible y se ha documentado de forma clara en el presente artículo.

Preprocesamiento de imágenes

Para adaptar estas imágenes monocromáticas a la entrada esperada por el modelo preentrenado *DenseNet121* (que requiere tres canales RGB), se replicó la matriz de escala de grises en los tres canales. Este enfoque permite preservar la integridad de la información radiológica mientras se asegura la compatibilidad estructural con modelos convolucionales entrenados en *ImageNet*.

Estrategia de validación cruzada estratificada

Con el propósito de obtener una evaluación robusta y generalizable del modelo, se aplicó una validación cruzada estratificada de 5 pliegues (Stratified K-Fold, $k = 5$), asegurando que la proporción de clases se mantuviera en cada partición. Esta técnica evita sesgos de evaluación comunes en conjuntos desbalanceados y proporciona métricas promedio más representativas del comportamiento real del modelo.

Balanceo de clases

Dado que el conjunto de datos presenta un desequilibrio importante entre las clases (PNEUMONIA es mayoritaria), se implementó un procedimiento de *oversampling* aleatorio sobre la clase minoritaria (NORMAL) dentro de cada pliegue de entrenamiento. Se utilizó la técnica de remuestreo con reemplazo para equilibrar ambas clases, lo que previno el sesgo hacia la clase dominante y mejoró la capacidad del modelo para detectar ambas condiciones clínicas.

Aumento de datos

Para incrementar la variabilidad de las muestras y reducir el sobreajuste, se aplicó *data augmentation* utilizando la clase *ImageDataGenerator* de Keras, transformaciones tales como la rotación, zoom, flip están fundamentadas en metodologías consolidadas en literatura [3]. Las transforma-

ciones incluyeron rotación aleatoria ($\pm 15^\circ$), desplazamientos horizontales y verticales ($\pm 10\%$), zoom aleatorio (hasta 20%) y volteo horizontal. Este paso fue realizado únicamente sobre los datos de entrenamiento, nunca sobre el conjunto de validación.

Arquitectura del modelo

Se adoptó la arquitectura *DenseNet121* como modelo base debido a su eficiencia en la reutilización y propagación directa de características entre capas [2], lo cual la ha posicionado como una de las arquitecturas más efectivas en tareas de clasificación médica. Se utilizó la versión preentrenada en *ImageNet* y se aplicó un *fine-tuning* parcial, todas las capas del modelo base hasta el índice 100 se mantuvieron congeladas (incluyendo las capas de convolución, normalización por lotes y activaciones correspondientes al bloque inicial y parte del primer bloque denso), mientras que las capas restantes permanecieron entrenables. Esta configuración permitió conservar las representaciones visuales generales aprendidas durante el preentrenamiento y adaptar únicamente las capas finales al dominio clínico, reduciendo el riesgo de sobreajuste. La cabeza original del clasificador fue reemplazada por una capa *Global Average Pooling*, seguida de *Dropout* (30 %), una capa densa de 128 neuronas con activación ReLU, un segundo *Dropout* del 30 %, y finalmente una capa de salida sigmoide para la clasificación binaria NORMAL y PNEUMONIA.

Entrenamiento y optimización

El modelo fue entrenado utilizando la función de pérdida binary crossentropy, ampliamente recomendada para tareas de clasificación binaria, y el optimizador Adam, respaldado por Kingma & Ba como método eficiente para aprendizaje estocástico [4], con una tasa de aprendizaje inicial de $1e-4$, debido a su eficacia comprobada en contextos de aprendizaje profundo [5]. Para prevenir el sobreajuste desde el diseño estructural de la red, se empleó Dropout, técnica propuesta por [6] para reducir la coadaptación de características [6].

Adicionalmente, se incorporaron dos estrategias dinámicas de regularización durante el proceso de entrenamiento. La primera fue *EarlyStopping*, que detuvo el aprendizaje cuando no se observó mejora en la pérdida de validación (*val_loss*) durante tres épocas consecutivas, evitando un entrenamiento excesivo. La segunda fue *ReduceLROnPlateau*, la cual redujo automáticamente la tasa de aprendizaje en un 30 % cuando la pérdida de validación no mejoró durante dos épocas, favoreciendo una convergencia más estable en etapas avanzadas del entrenamiento. Cada pliegue fue entrenado durante un máximo de 20 épocas, con un tamaño de *batch* de 32 imágenes. La validación se realizó utilizando el subconjunto correspondiente al *fold*, asegurando una evaluación honesta y libre de fugas de información.

Evaluación del modelo

La evaluación del modelo se llevó a cabo al finalizar cada pliegue de la validación cruzada, empleando un conjunto integral de métricas para caracterizar su desempeño desde múltiples perspectivas. La métrica principal fue el AUC-ROC (Área bajo la curva receptor-operador), utilizada como indicador robusto de la capacidad del modelo para discriminar entre las clases NORMAL y PNEUMONIA, independientemente del umbral de decisión. Asimismo, se calcularon las métricas de precisión, recuperación (Recall) y F1-Score para cada clase, mediante el método *classification_report()* de la biblioteca Scikit-learn [7], con el fin de proporcionar una visión deta-

llada de la calidad de las predicciones en términos de relevancia y cobertura. Posteriormente se construyeron matrices de confusión por pliegue, que permitieron visualizar de manera explícita los patrones de clasificación correcta y los errores sistemáticos.

RESULTADOS

El modelo propuesto, basado en la arquitectura *DenseNet121* con ajuste fino parcial y validación cruzada estratificada de cinco pliegues, mostró un rendimiento altamente robusto y consistente en la tarea de clasificación binaria de imágenes de tórax, diferenciando entre casos normales y neumonía. Durante cada pliegue, se registraron métricas individuales que posteriormente fueron promediadas para proporcionar una estimación generalizada del desempeño del sistema.

En cuanto a la métrica *AUC-ROC*, considerada uno de los indicadores más confiables en contextos clínicos por su independencia del umbral de clasificación, se alcanzó un valor promedio de 0,9967, lo que evidencia una capacidad casi perfecta del modelo para discriminar entre pacientes sanos y aquellos con neumonía. Esta métrica se mantuvo estable a lo largo de los cinco pliegues, con valores individuales que oscilaron entre 0,9960 y 0,9981, lo que refuerza la robustez del enfoque incluso frente a variaciones leves en la distribución de los datos.

En términos de precisión, recuperación y *F1-score*, se obtuvieron resultados destacables para ambas clases. La clase PNEUMONIA presentó una precisión promedio de 0,999, una recuperación de 0,888 y un *F1-score* de 0,939, reflejando un excelente balance entre sensibilidad y especificidad, crucial en aplicaciones médicas donde el costo de los falsos negativos es crítico. La clase NORMAL, por su parte, alcanzó una precisión promedio de 0,761, una recuperación de 0,996 y un *F1-score* de 0,861, lo que sugiere que el modelo es altamente competente en identificar correctamente a los pacientes sanos, aunque presenta todavía casos de neumonía clasificados como normales, de esta manera se buscó mantener una precisión muy alta en la clase PNEUMONIA y no incrementar en exceso las predicciones erróneas.

El análisis de las matrices de confusión mostró que, pese a la alta capacidad discriminativa global, todavía se observan falsos negativos casos de neumonía clasificados como normales, consistentes con un *Recall* de 0,888 en la clase PNEUMONIA. Esto indica que el modelo, aunque muy preciso en sus predicciones positivas $\text{precision} \approx 0,99$, aún puede pasar por alto ciertos casos patológicos, lo que debe considerarse con cautela en escenarios de tamizaje clínico.

Los resultados obtenidos demuestran que la arquitectura *DenseNet121* combinada con un esquema de validación cruzada estratificada y técnica de balanceo y aumento de datos, constituye una solución precisa y metodológicamente robusta a nivel de validación interna, con potencial para ser clínicamente útil, aunque requiere validación externa antes de considerarse plenamente generalizable.

Fuente: Elaboración propia.

Comparación de Métricas de Validación Cruzada por Fold

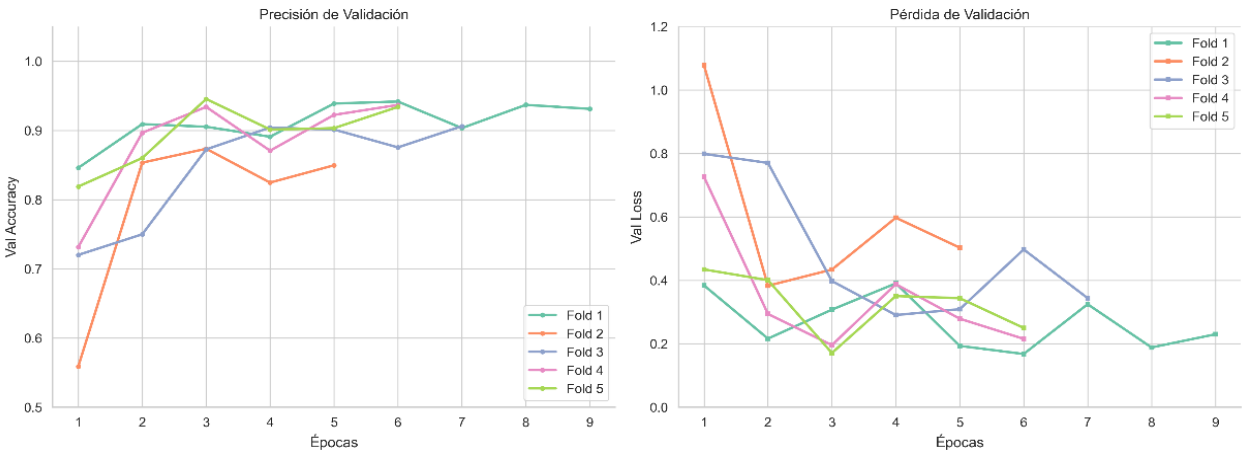


Figura 1. Comparación de métricas de validación cruzada por Fold.

Tabla 1. Evolución del modelo durante entrenamiento y validación en 5-Fold cross validation.

Fold	Val Accuracy	Val Loss	Train Accuracy	Train Loss
Fold 1	0,931	0,2298	0,9879	0,0295
Fold 2	0,8495	0,5028	0,9785	0,0659
Fold 3	0,9060	0,3439	0,9830	0,0503
Fold 4	0,9367	0,2154	0,9783	0,0539
Fold 5	0,9338	0,2498	0,9732	0,0647

Fuente: Elaboración propia.

La [Tabla 1](#) muestra los resultados de la última época entrenada por *Fold* obtenidos mediante validación cruzada, se muestran la precisión y pérdida tanto en entrenamiento como en validación, donde se observa un alto rendimiento en el *Fold 4* con 0,9367 de precisión y 0,2154 de pérdida.

Tabla 2. AUC-ROC y F1 promedio por clase en validación.

Fold	AUC-ROC	NORMAL		PNEUMONIA	
		F1	Recall	F1	Recall
Fold 1	0,9981	0,8982	0,9998	0,9590	0,9213
Fold 2	0,9962	0,7773	0,9963	0,8906	0,8039
Fold 3	0,9964	0,8423	0,9963	0,9311	0,8723
Fold 4	0,9965	0,8852	0,9925	0,9535	0,9135
Fold 5	0,9960	0,9036	0,9963	0,9619	0,9277
Media	0,9967	0,861	0,996	0,939	0,888

Fuente: Elaboración propia.

La [Tabla 2](#) muestra que el modelo mantiene un rendimiento altamente consistente durante la validación cruzada estratificada de cinco pliegues, con valores de AUC-ROC entre 0,9960 y 0,9981

media 0,9967, lo que confirma una excelente capacidad de discriminación entre radiografías normales y con neumonía. El desempeño promedio del modelo fue superior en la clase PNEUMONIA $F1 = 0,939$ respecto a la clase NORMAL $F1 = 0,861$, lo que refleja una combinación de precisión muy alta $\approx 0,99$ y una sensibilidad todavía mejorable $\approx 0,888$ en los casos de neumonía, lo cual es clave a considerar en aplicaciones clínicas.

Tabla 3. Comparación del rendimiento del modelo propuesto frente a estudios previos.

Estudio	Modelo	Protocolo	AUC-ROC	F1 Normal	F1 neumonía
CheXNet [1]	DenseNet121	Train/test (hospital-scale)	0,9		0,435
Chest X-ray Pneumonia - CNN Transfer Learning [23]	CNN con transfer learning	Train/val/test en hold-out	0,978	0,88	0,94
Trabajo propuesto en el presente artículo	DenseNet121 con finetuning	5-fold cross-validation estratificada (validación)	0,9967	0,861	0,939

Fuente: Elaboración propia.

La [Tabla 3](#) muestra que el modelo propuesto alcanza el mejor rendimiento global, con un AUC-ROC promedio de 0,9967 y un F1 elevado en la clase neumonía (0,939), superando la capacidad discriminativa reportada en *CheXNet* [1] y en la implementación pública de Kaggle [23]. Además, a diferencia del enfoque en *hold-out* utilizado en ese trabajo, nuestro modelo fue evaluado mediante validación cruzada estratificada, lo que aporta mayor robustez a los resultados y reduce la dependencia de una sola partición de los datos. Estos hallazgos confirman que la propuesta mejora el desempeño y estabilidad de métodos previos dentro del mismo dominio diagnóstico.

Fuente: Elaboración propia.

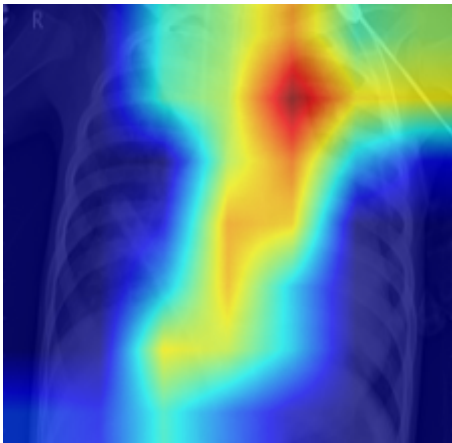


Figura 2. Visualización Grad-CAM de un caso positivo de neumonía.

La [Figura 2](#) muestra el mapa de activación *Grad-CAM* superpuesto sobre una radiografía de tórax clasificada correctamente como neumonía por el modelo propuesto. Las regiones en color rojo y amarillo reflejan las áreas con mayor contribución a la predicción, ubicadas principalmente en el parénquima pulmonar central e inferior, donde se observan opacidades compatibles con consolidación. Este patrón de atención indica que el modelo se enfoca en estructuras clínicas relevantes para el diagnóstico, lo cual contribuye a su interpretabilidad en entornos médicos y sugiere que la

red no está aprendiendo de artefactos externos como etiquetas, bordes o dispositivos médicos.

La visualización mediante *Grad-CAM* muestra que el modelo focaliza su atención principalmente en el parénquima pulmonar y en las opacidades de la región central e inferior del tórax, lo cual es coherente con patrones radiológicos típicos de neumonía. Esto sugiere que el modelo aprende características clínicas relevantes y no depende de artefactos externos como textos, bordes o cables presentes en la imagen, de esta manera demostramos que nuestro modelo apoya también en la interpretabilidad.

DISCUSIÓN

Los resultados obtenidos en este estudio demuestran que es posible alcanzar niveles muy altos de precisión y capacidad discriminativa en la detección de neumonía utilizando redes convolucionales profundas como *DenseNet121*, incluso partiendo de imágenes médicas en escala de grises y sin requerir una infraestructura de datos clínicos complejos. El promedio de AUC-ROC alcanzado 0,9967 representa un desempeño que compite con estudios de referencia en la literatura, y se obtuvo en un entorno de validación cruzada estratificada, lo cual fortalece la confiabilidad del modelo frente a variaciones en la distribución de los datos.

Una de las decisiones metodológicas más relevantes fue el uso de técnicas de aumento de datos combinadas con un balanceo activo mediante sobre-muestreo de la clase minoritaria. Estas estrategias permitieron compensar parcialmente el desequilibrio de clases, uno de los problemas más frecuentes en *datasets* médicos. El uso de validación cruzada no solo permitió evaluar la estabilidad del modelo en distintos subconjuntos de datos, sino que también reveló una consistencia destacable en las métricas obtenidas entre pliegues, especialmente en la clase asociada a la condición patológica.

Si bien existen soluciones previas sobre el *dataset*, nuestro estudio aporta una metodología más rigurosa basada en validación cruzada estratificada con múltiples métricas, por ejemplo, en [23] desarrollaron un modelo de transferencia de aprendizaje para la clasificación binaria de neumonía pediátrica en radiografías de tórax, alcanzando un AUC-ROC de 0,978 y una exactitud de 0,92 en el conjunto de prueba. Sin embargo, dicho trabajo se basó en una única partición entrenamiento-prueba, lo que puede incrementar el riesgo de sobreajuste y limita la estimación real de la capacidad de generalización del modelo. Por otro lado, el enfoque propuesto en este estudio emplea una validación cruzada estratificada de 5 pliegues, lo que permite obtener métricas más estables y representativas. Los resultados muestran una mejora significativa en el desempeño con un AUC-ROC promedio de 0,9967 y un rendimiento consistente entre pliegues, reforzando la robustez metodológica y la contribución del modelo frente a implementaciones previas de acceso público.

Los autores Kundu, Das, Geem, Han y Sarkar [11], propusieron un *ensemble* de *DenseNet-121*, *ResNet-18* y *GoogleNet* combinado por métricas múltiples, alcanzando 98,8% de *accuracy* con validación cruzada estratificada en la base de datos de Kermany [11]. Los autores Rashid, Asma, Barua y Das [24], mejoraron *DenseNet-121* integrando módulos CBAM, elevando la exactitud del 95,6% al 98,6% y utilizando *Grad-CAM* para resaltar las áreas pulmonares afectadas, lo que evidenció el impacto positivo de los mecanismos de atención en esta arquitectura *DenseNet-121*.

[24].

Por otro lado, emplearon una arquitectura híbrida *DenseNet + Vit* [25], logrando 97,9% de precisión en un escenario *multi-clase* y validando con 5-fold *cross-validation*. El uso de auto-atención global y evaluación contrasta con el modelo del presente trabajo poniendo en valor la integración de *Transformers* y técnicas de validación estrictas. Así como también, modificaron *DenseNet121* con bloques *Squeeze and Excitation* [26], de esa manera obtuvieron un pequeño incremento en la precisión de 90% a 93% aproximadamente y *recall* en detección de neumonía, esto nos enseña que inclusive mecanismos de atención simples como se pueden aportar mejoras sobre *DenseNet*. En [27] emplearon transferencia de aprendizaje con múltiples CNN y un módulo de atención para fusionar sus características, obteniendo 96,6% de precisión en el diagnóstico de neumonía, donde resaltaron la importancia y ventaja de combinar extractores diferentes y atención para potenciar la exactitud al diagnosticar.

En comparación con otros enfoques existentes, que muchas veces reportan métricas elevadas, pero en esquemas de validación poco robustos (por ejemplo, sin estratificación o sin desbalance controlado), este trabajo refuerza la importancia de una evaluación metodológicamente rigurosa para evitar sobreestimaciones de rendimiento. Incluso con una arquitectura preentrenada, se evidenció que un *fine-tuning* parcial bien diseñado puede mejorar sustancialmente la adaptación del modelo a un nuevo dominio de datos, sin necesidad de un entrenamiento completo desde cero.

Cabe destacar, no obstante, que el presente estudio trabajó exclusivamente con datos del conjunto *chest_xray*, lo cual limita de alguna forma la capacidad de generalización externa. Si bien los resultados son notables, no es posible asumir que el modelo tendrá el mismo comportamiento sobre datos adquiridos en hospitales con protocolos de imagen distintos, diferentes poblaciones o equipos de menor resolución.

El *recall* cercano al 100% en la clase NORMAL, combinado con una menor sensibilidad en la clase PNEUMONIA ($\text{recall} \approx 0,888$), sugiere que el modelo tiende a ser conservador a la hora de etiquetar como neumonía, lo que puede derivar en algunos casos patológicos no detectados. Este patrón es aceptable solo si el sistema se usa como apoyo complementario y siempre bajo supervisión médica.

Este trabajo representa un paso hacia modelos que no solo sean precisos, sino también realistas y replicables. Integrar explicación precisa, validación multicéntrica y retroalimentación médica continua serán elementos clave en siguientes etapas para que este tipo de soluciones se adopten efectivamente en la práctica clínica.

Es importante aclarar que el sistema propuesto no pretende sustituir la evaluación médica. El modelo se plantea como una herramienta de apoyo a la toma de decisiones, que tenga la capacidad de priorizar casos sospechosos de neumonía para agilizar el proceso diagnóstico en entornos con recursos limitados. Su uso debe realizarse en conjunto con la valoración clínica del paciente y la interpretación experta de un profesional de salud, dado que las radiografías por sí solas no constituyen un estándar diagnóstico definitivo.

Es importante señalar que el presente estudio constituye una etapa metodológica preliminar dentro de una investigación más amplia actualmente en desarrollo, en la siguiente fase del proyecto correspondiente a una tesis, el modelo será ajustado y validado utilizando radiografías de tórax de la región. Esta evaluación local permitirá analizar el rendimiento del sistema en un entorno clínico específico caracterizado por limitaciones de recursos y menor disponibilidad de radiólogos.

Variabilidad estadística y justificación muestral

Para evaluar la estabilidad del rendimiento del modelo, se utilizó validación cruzada estratificada de cinco pliegues, lo que permitió obtener métricas promedio junto con su variabilidad. Los valores de AUC-ROC oscilaron entre 0,9960 y 0,9981, evidenciando una baja dispersión y, por tanto, alta consistencia del modelo. El tamaño muestral utilizado corresponde al conjunto de datos *Chest X-ray Images Pneumonia*, ampliamente empleado y validado en estudios de detección de neumonía pediátrica a nivel internacional, lo que respalda su idoneidad para fines comparativos y metodológicos.

Implicancias clínicas

Los resultados obtenidos sugieren que el modelo puede integrarse como una herramienta de apoyo para los profesionales de la salud, como una segunda opinión de diagnóstico rápido y consistente. Su implementación no reemplaza el criterio médico, sino que contribuye a priorizar casos y reducir la carga operativa en entornos con recursos limitados.

Líneas futuras de trabajo

Se plantea como trabajo futuro validar el modelo con radiografías locales y ampliar el conjunto de datos hacia diferentes condiciones pulmonares. Asimismo, se propone incorporar información clínica complementaria y evaluar el rendimiento en escenarios hospitalarios reales para una adopción más segura y efectiva. Además, se recomienda experimentar con algoritmos optimizadores cuánticos.

CONCLUSIONES

Los hallazgos obtenidos a lo largo del estudio ponen en evidencia la eficacia de una arquitectura como *DenseNet121*, cuando es adecuadamente adaptada a un problema clínico específico como la detección de neumonía en imágenes de rayos X torácicas. El uso de técnicas de validación cruzada estratificada, junto con un cuidadoso preprocesamiento, balanceo de clases y regularización, permitió construir un sistema de clasificación binaria con un rendimiento altamente consistente.

El valor promedio de AUC-ROC obtenido, superior al 0,996, sugiere una capacidad de discriminación excepcional, difícilmente atribuible al azar o al sobreajuste. Además, la estabilidad de las métricas obtenidas en cada pliegue del experimento demuestra que el modelo generaliza bien sobre subconjuntos del conjunto de entrenamiento, lo que fortalece su validez para aplicaciones futuras en contextos reales.

Uno de los aspectos más destacables es la forma en que se abordaron las limitaciones típicas de estos *datasets*, como el desequilibrio entre clases y la ausencia de etiquetas múltiples o gradaciones de severidad. Las estrategias empleadas no solo permitieron mitigar estos problemas, sino que lo hicieron sin comprometer el poder predictivo del modelo, un balance que suele ser complejo de lograr.

Si bien los resultados son alentadores, todavía existen aspectos que pueden perfeccionarse. Aunque en este trabajo se incorporó una visualización preliminar mediante *Grad-CAM*, la integración sistemática de herramientas interpretables y su evaluación junto a especialistas clínicos sería un paso importante hacia su adopción clínica. Además, evaluar su comportamiento con imágenes provenientes de otras regiones geográficas o sistemas de adquisición más rudimentarios permitiría comprender su robustez en casos menos controlados.

Más allá de las métricas, este trabajo sugiere que, con el tratamiento metodológico adecuado, los modelos de visión por computadora pueden convertirse en aliados reales para el personal médico, especialmente en regiones con limitaciones tecnológicas o recursos humanos escasos.

AGRADECIMIENTOS

Los autores expresan su sincero agradecimiento al Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos (IIICCD) por el acompañamiento técnico y la orientación brindada durante el desarrollo de este estudio, así como por facilitar el acceso a recursos de cómputo y análisis especializado. Se reconoce también el apoyo de la Facultad de Ingeniería Estadística e Informática por fomentar la investigación científica en áreas interdisciplinarias y por su respaldo institucional. A los docentes y asesores que, mediante sus conocimientos y comentarios críticos, contribuyeron al fortalecimiento metodológico del trabajo, nuestro más profundo reconocimiento. Finalmente, se agradece a la Universidad Nacional del Altiplano de Puno, por proporcionar el entorno académico propicio que permitió la realización de esta investigación.

Financiamiento

Los autores declaran no haber recibido financiamiento para este trabajo.

REFERENCIAS

- [1] P. Rajpurkar et al., “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” arXiv:1711.05225, 2017. [†]
- [2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261-2269, <https://doi.org/10.1109/CVPR.2017.243> [†]
- [3] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, Art. no. 60, 2019, <https://doi.org/10.1186/s40537-019-0197-0> [†]

- [4] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization, ” arXiv:1412.6980, 2014. [†]
- [5] F. Chollet, *Deep Learning with Python*, 1st ed. Shelter Island, New York, USA: Manning Publications, 2017. [†]
- [6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting, ” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [†]
- [7] F. Pedregosa et al., “Scikit-learn: Machine learning in Python, ” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [†]
- [8] G. Litjens et al., “A survey on deep learning in medical image analysis, ” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, <https://doi.org/10.1016/j.media.2017.07.005> [†]
- [9] S. Sukhendra Singh, M. Kumar, A. Kumar, B. Kumar Verma, K. Abhishek, and S. Selvarajan, “Efficient pneumonia detection using Vision Transformers on chest X-rays, ” *Scientific Reports*, vol. 14, Art. 2487, 2024, <https://doi.org/10.1038/s41598-024-52703-2> [†]
- [10] A. Musa, R. Prasad, and M. Hernandez, “Addressing cross-population domain shift in chest X-ray classification through supervised adversarial domain adaptation, ” *Scientific Reports*, vol. 15, Art. 11383, 2025, <https://doi.org/10.1038/s41598-025-95390-3> [†]
- [11] R. Kundu, R. Das, Z. W. Geem, G. T. Han, and R. Sarkar, “Pneumonia detection in chest X-ray images using an ensemble of deep learning models, ” *Plos One*, vol. 16, no. 9, Art. no. e0256630, 2021, <https://doi.org/10.1371/journal.pone.0256630> [†]
- [12] J. Hofmeister et al., “Validating the accuracy of deep learning for the diagnosis of pneumonia on chest X-ray against a robust multimodal reference diagnosis: A post hoc analysis of two prospective studies, ” *European Radiology Experimental*, vol. 8, no. 1, Art. no. 20, 2024, <https://doi.org/10.1186/s41747-023-00416-y> [†]
- [13] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, “An efficient deep learning approach to pneumonia classification in healthcare, ” *Journal of Healthcare Engineering*, vol. 2019, Art. No. 4180949, 2019, <https://doi.org/10.1155/2019/4180949> [†]
- [14] A. E. W. Johnson et al., “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, ” *Scientific Data*, vol. 6, no. 1, Art. no. 317, 2019, <https://doi.org/10.1038/s41597-019-0322-0> [†]
- [15] P. Rajpurkar et al., “Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, ” *PLoS Medicine*, vol. 15, no. 11, Art. no. e1002686, 2018, <https://doi.org/10.1371/journal.pmed.1002686> [†]
- [16] J. S. Ahn et al., “Association of artificial intelligence–aided chest radiograph interpretation with reader performance and efficiency, ” *JAMA Network Open*, vol. 5, no. 8, Art. no. e2229289, 2022, <https://doi.org/10.1001/jamanetworkopen.2022.29289> [†]
- [17] J. C. Y. Seah et al., “Effect of a comprehensive deep-learning model on the accuracy of chest X-ray interpretation by radiologists: A retrospective, multireader multicase study, ” *The Lancet Digital Health*, vol. 3, no. 8, pp. e496–e506, 2021, [https://doi.org/10.1016/S2589-7500\(21\)00106-0](https://doi.org/10.1016/S2589-7500(21)00106-0) [†]

- [18] D. S. Kermany, K. Zhang, and M. Goldbaum, January 6, 2018 “Labeled optical coherence tomography (OCT) and chest X-ray images for classification, ” Mendeley Data, V2, <https://doi.org/10.17632/rscbjbr9sj.2> [↑]
- [19] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and benchmarks on weakly-supervised classification and localization of common thorax diseases” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 3462-3471, <https://doi.org/10.1109/CVPR.2017.369> [↑]
- [20] J. Irvin et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, ” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no.1, pp. 590–597, 2019, <https://doi.org/10.1609/aaai.v33i01.3301590> [↑]
- [21] G. Liang and L. Zheng, “A transfer learning method with deep residual network for pediatric pneumonia diagnosis, ” *Computer Methods and Programs in Biomedicine*, vol. 187, Art. no. 104964, 2020, <https://doi.org/10.1016/j.cmpb.2019.06.023> [↑]
- [22] M. Bunea and G. M. Danciu, “Pneumonia image classification using DenseNet architecture, ” *Information*, vol. 15, no. 10, Art. no. 611, 2024, <https://doi.org/10.3390/info15100611> [↑]
- [23] J. P. Barbosa, “Chest X-Ray (Pneumonia) - CNN & Transfer Learning, ” Kaggle.com. [Online]. Available: <https://www.kaggle.com/code/jonaspalucibarbosa/chest-x-ray-pneumonia-cnn-transfer-learning> [↑]
- [24] A. B. Rashid, J. Asma, K. Barua, and D. Das, “An enhanced deep learning framework for pneumonia detection in Chest X-rays, ” *SN Computer Science*, vol. 6, no. 5, Art. no. 472, 2025, <https://doi.org/10.1007/s42979-025-04017-x> [↑]
- [25] B. Oltu, S. Güney, S. E. Yuksel, and B. Dengiz, “Automated classification of chest X-rays: a deep learning approach with attention mechanisms, ” *BMC Medical Imaging*, vol. 25, Art. no. 71, 2025. [↑]
- [26] K. Wang, P. Jiang, J. Meng, and X. Jiang, “Attention-based DenseNet for pneumonia classification, ” *IRBM*, vol. 43, no. 5, pp. 479–485, 2022. [↑]
- [27] S.-M. Cha, S.-S. Lee, and B. Ko, “Attention-based transfer learning for efficient pneumonia detection in Chest X-ray images, ” *Applied Sciences*, vol. 11, no. 3, Art. 1242, 2021, <https://doi.org/10.3390/app11031242> [↑]