SaDe-REVISTA DE CIENCIAS DE LA SALUD Y EL DEPORTE
AÑO IX | N° 6 | JULIO DE 2026

Inteligencia artificial reproducible en educación superior

Desarrollo y despliegue de un modelo de detección de neumonía con aprendizaje transferido

*Leandro Guerschberg, Lorena Rivero y Yael Estefanía Gutierrez**

Resumen

Este artículo presenta un estudio orientado al desarrollo de un sistema de clasificación automática de radiografías de tórax para diferenciar imágenes normales de casos con neumonía, empleando técnicas contemporáneas de aprendizaje profundo. Se utilizó un conjunto de datos público ampliamente difundido en investigación biomédica y se realizó un proceso de *fine-tuning* sobre la arquitectura ResNet18 previamente entrenada en ImageNet. El entrenamiento se llevó a cabo en la plataforma Kaggle, optimizando el rendimiento mediante estrategias de transferencia de aprendizaje y selección del mejor modelo según desempeño en validación. El sistema alcanzó un 89,4% de *accuracy* (precisión) en el conjunto de prueba, con elevadas métricas de sensibilidad para la detección de neumonía. Posteriormente, se desarrolló una aplicación web interactiva mediante Gradio y se desplegó en HuggingFace Spaces, permitiendo a usuarios cargar radiografías y obtener predicciones acompa-

* Leandro Guerschberg, Departamento de Ciencias de la Salud y Deporte (DCSyD), Universidad Nacional de José C. Paz, Provincia de Buenos Aires, Argentina. Contacto: Leandro.guerschberg@docentes.unpaz.edu.ar ORCID: 0009-0005-9286-6358
Lorena Rivero, Departamento de Ciencias de la Salud y Deporte (DCSyD), Universidad Nacional de José C. Paz, Provincia de Buenos Aires, Argentina. Contacto: lrivero@unpaz.edu.ar ORCID: 0009-0009-4727-167X
Yael Estefanía Gutierrez, Departamento de Ciencias de la Salud y Deporte (DCSyD), Universidad Nacional de José C. Paz, Provincia de Buenos Aires, Argentina. Contacto: yael.gutierrez@docentes.unpaz.edu.ar ORCID: 0009-0003-1616-2995cional de José C. Paz. Administración Nacional de Laboratorios y Salud (ANLIS), Instituto Nacional de Parasitología "Dr. Mario Fatala Chaben".

ñadas de probabilidades. El trabajo propone un flujo de trabajo completamente reproducible con herramientas gratuitas, aplicable tanto a la enseñanza universitaria en informática y ciencias de la salud como a proyectos de investigación formativos en imágenes médicas.

Palabras clave: inteligencia artificial - neumonía - aprendizaje transferido - ResNet18 - educación superior

Abstract

This article presents the development of an automated chest X-ray classification system designed to distinguish normal images from cases of pneumonia using contemporary deep learning techniques. A widely used public biomedical dataset was employed, and transfer learning was implemented through fine-tuning of a ResNet18 architecture pretrained on ImageNet. Model training was performed on the Kaggle platform, optimizing performance through transfer learning strategies and selecting the best-performing model based on validation results. The final system achieved an accuracy of 89.4% on the test dataset while demonstrating high sensitivity for pneumonia detection. Subsequently, an interactive web application was developed using Gradio and deployed on Hugging Face Spaces, enabling users to upload chest X-ray images and receive diagnostic predictions together with probability estimates. The study proposes a fully reproducible workflow based on freely available tools that can be applied in higher education for teaching computer science and health sciences, as well as in educational research projects involving medical imaging.

Keywords: artificial intelligence - pneumonia - transfer learning - higher education

Introducción

La interpretación de radiografías de tórax constituye una de las prácticas diagnósticas más extendidas en la medicina clínica, debido a su accesibilidad, bajo costo relativo y relevancia para la identificación de patologías prevalentes. Entre estas, la neumonía continúa siendo una de las principales causas de morbimortalidad a nivel global, particularmente en contextos de alta demanda asistencial o con limitada disponibilidad de especialistas. En este escenario, el desarrollo de herramientas basadas en inteligencia artificial (IA) orientadas a la automatización parcial del análisis radiográfico ha adquirido creciente interés, tanto para el apoyo clínico como para su incorporación en ámbitos educativos y de investigación.

El avance de las redes neuronales convolucionales (CNN) ha permitido mejorar significativamente el procesamiento automático de imágenes médicas. En particular, la arquitectura ResNet, introducida por He, Zhang, Ren y Sun (2016), marcó un punto de inflexión al permitir el entrenamiento eficiente

de redes profundas mediante el uso de conexiones residuales. Este desarrollo, junto con la disponibilidad de grandes repositorios de imágenes radiológicas –como ChestX-ray14 (Wang et al, 2017), CheXpert (Irvin et al, 2019) y CheXNet (Rajpurkar et al, 2017)–, consolidó un marco experimental robusto para la aplicación de aprendizaje profundo en diagnóstico por imágenes.

Paralelamente, la emergencia de *frameworks* de desarrollo como PyTorch (Paszke et al, 2019) y TensorFlow (Abadi et al, 2015), junto con plataformas de cómputo en la nube como Kaggle, ha contribuido a democratizar el acceso a herramientas de entrenamiento y despliegue de modelos, permitiendo que instituciones educativas sin infraestructura especializada puedan reproducir experiencias de aprendizaje profundo en contextos formativos. Este escenario se articula con la noción de "medicina de alto rendimiento" propuesta por Topol (2019), que plantea una convergencia entre capacidades humanas y sistemas de IA orientada a mejorar la precisión diagnóstica y optimizar los procesos clínicos, sin reemplazar el juicio profesional.

No obstante, a pesar de estos avances, persisten barreras significativas para la incorporación efectiva de la IA en la formación de profesionales de la salud. Muchos de los modelos desarrollados en la literatura presentan altos niveles de complejidad, requieren recursos computacionales elevados o se encuentran integrados en plataformas propietarias, lo que dificulta su utilización en entornos educativos. En consecuencia, estudiantes y docentes enfrentan limitaciones para experimentar con datos reales, comprender los procesos de entrenamiento y analizar críticamente el funcionamiento de estos sistemas.

En este contexto, la presente investigación se propone abordar estas limitaciones mediante el desarrollo de un modelo de clasificación de radiografías de tórax basado en la arquitectura ResNet18, seleccionada por su equilibrio entre capacidad de representación y accesibilidad computacional. A través de la implementación de técnicas de transferencia de aprendizaje y el uso exclusivo de herramientas gratuitas, el estudio busca construir un flujo de trabajo reproducible que abarque desde el entrenamiento del modelo hasta su despliegue en una aplicación web interactiva.

La relevancia de este enfoque radica en su potencial para ser replicado en entornos de educación superior, particularmente en carreras vinculadas a las ciencias de la salud, donde la integración de la inteligencia artificial requiere no solo resultados técnicos, sino también propuestas pedagógicas que permitan comprender sus alcances, limitaciones y desafíos éticos. En este sentido, el desarrollo de herramientas accesibles y reproducibles se presenta como una estrategia clave para promover la alfabetización digital crítica y la apropiación significativa de tecnologías emergentes en contextos formativos.

Objetivo general

Desarrollar, entrenar, evaluar y desplegar un modelo de clasificación automática de radiografías de tórax para distinguir imágenes normales de casos con neumonía mediante técnicas de aprendizaje

profundo y herramientas gratuitas, con fines de apoyo a la enseñanza y la investigación formativa en ciencias de la salud.

Objetivos específicos

1. *Comprender* los fundamentos conceptuales del aprendizaje profundo y su aplicación en la clasificación de imágenes médicas, revisando la literatura actual sobre arquitecturas convolucionales, *datasets* radiológicos y modelos de apoyo al diagnóstico.
2. *Aplicar* técnicas de *transfer learning* mediante la adaptación de ResNet18 para la clasificación binaria de radiografías de tórax utilizando un conjunto de datos público y plataformas de entrenamiento accesibles.
3. *Analizar* el rendimiento del modelo entrenado a través de métricas estandarizadas –*accuracy*, precisión, *recall*, F1-*score* y matriz de confusión– interpretando sus implicancias diagnósticas y educativas.
4. *Evaluar* las limitaciones, fortalezas y sesgos potenciales del sistema, considerando elementos clínicos, técnicos y éticos propios del uso de IA en salud.
5. *Crear* una aplicación web interactiva que permita cargar nuevas imágenes radiográficas y obtener predicciones acompañadas de probabilidades, utilizando herramientas abiertas como Gradio y HuggingFace Spaces, con el propósito de facilitar la apropiación pedagógica del modelo por parte de docentes y estudiantes.

Materiales y métodos

El presente estudio se desarrolló bajo un enfoque experimental orientado a la reproducibilidad metodológica y a la accesibilidad en contextos educativos. El diseño se estructuró en siete dimensiones: selección del *dataset*, preprocesamiento de imágenes, definición de la arquitectura del modelo, configuración del entrenamiento, evaluación cuantitativa, desarrollo de la interfaz de usuario y despliegue en un entorno abierto.

Dataset y fuentes de datos

Se utilizó el *dataset* "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification" (Kermany, Zhang y Goldbaum, 2018), disponible en Mendeley Data. Este conjunto incluye radiografías de tórax pediátricas clasificadas en dos categorías: *NORMAL* y *PNEUMONIA*, organizadas en subconjuntos de entrenamiento, validación y prueba. Su uso resulta pertinente por su accesibilidad y por presentar variabilidad en condiciones de adquisición, niveles de exposición y manifestaciones clínicas, lo que favorece el análisis de la capacidad de generalización del modelo.

Si bien existen repositorios de mayor escala como ChestX-ray14 (Wang et al, 2017) o CheXpert (Irvin et al, 2019), su complejidad y volumen exceden los requerimientos de un enfoque formativo inicial.

En este sentido, el *dataset* seleccionado ofrece un equilibrio adecuado entre realismo clínico y manejabilidad computacional.

Desbalance de clases

El *dataset* presenta un desbalance significativo en el conjunto de entrenamiento (≈25,7% *NORMAL* y ≈74,3% *PNEUMONIA*), lo que puede inducir sesgos en el aprendizaje. Este tipo de distribución suele favorecer la sensibilidad para la clase mayoritaria en detrimento de la identificación de la clase minoritaria. En el presente estudio, este comportamiento se refleja en un mayor *recall* para la clase *PNEUMONIA* y una reducción relativa en la clase *NORMAL*.

La mitigación del desbalance se abordó de manera indirecta mediante técnicas de *data augmentation*, que mejoran la capacidad de generalización sin modificar la distribución de clases. Asimismo, el uso de un conjunto de validación balanceado permitió una evaluación más equitativa durante el ajuste del modelo.

Preprocesamiento de imágenes

Las imágenes fueron redimensionadas a 224×224 píxeles para adecuarse a los requerimientos de la arquitectura ResNet18 y normalizadas utilizando los parámetros del *dataset* ImageNet (He et al, 2016), sobre el cual el modelo base fue preentrenado.

Durante el entrenamiento se aplicaron técnicas de *data augmentation* –incluyendo rotaciones leves, inversión horizontal y ajustes de brillo y contraste– con el objetivo de mejorar la robustez del modelo frente a variaciones propias de entornos clínicos reales. El *pipeline* de procesamiento se implementó en PyTorch, garantizando su reproducibilidad (Paszke et al, 2019).

Arquitectura del modelo

Se seleccionó la arquitectura ResNet18 (He et al, 2016), basada en aprendizaje residual mediante *skip connections*, que permiten un entrenamiento estable en redes profundas. Su elección responde a un equilibrio entre capacidad de representación, eficiencia computacional y adecuación a contextos educativos.

Se utilizó una versión preentrenada en ImageNet, reemplazando la capa final por un clasificador binario. Inicialmente se congelaron las capas convolucionales y se entrenó únicamente la capa final, siguiendo un enfoque de *transfer learning* que reduce los requerimientos de datos y acelera la convergencia.

Plataforma de entrenamiento y entorno experimental

El entrenamiento se realizó en la plataforma Kaggle, utilizando GPU Tesla T4 de acceso gratuito. El proceso se implementó mediante notebooks estructurados que incluyeron carga de datos, definición del modelo y ejecución del ciclo de entrenamiento.

Se utilizó el optimizador Adam con tasa de aprendizaje de 0,001, función de pérdida CrossEntropyLoss, tamaño de lote de 32 y un total de ocho épocas. Esta configuración prioriza la reproducibilidad y la estabilidad del entrenamiento en un entorno accesible.

Evaluación del modelo

La evaluación se llevó a cabo sobre el conjunto de prueba, utilizando métricas estándar en clasificación binaria: *accuracy*, precisión, *recall* y F1-*score*. En particular, el *recall* para la clase *PNEUMONIA* adquiere relevancia clínica, dado que minimiza la probabilidad de falsos negativos (Rajpurkar et al, 2017).

Se construyó además una matriz de confusión que permitió analizar el desempeño en términos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, constituyendo un recurso clave para la interpretación técnica y pedagógica del modelo.

Aunque no se incorporaron en esta etapa, se reconoce la relevancia de técnicas de explicabilidad como Grad-CAM (Selvaraju et al, 2017), que serán consideradas en desarrollos futuros.

Desarrollo de la aplicación web

Se desarrolló una aplicación web utilizando el *framework* Gradio (Gradio Team, 2023), que permite integrar modelos de aprendizaje profundo en interfaces interactivas. La aplicación incluye un cargador de imágenes, un módulo de inferencia y una visualización de resultados con la clase predicha y su probabilidad asociada.

El modelo se integró utilizando su versión exportada (.pth), replicando el mismo *pipeline* de preprocesamiento empleado durante el entrenamiento, lo que asegura consistencia en las predicciones.

Despliegue en entorno abierto

El sistema fue desplegado en HuggingFace Spaces, lo que permite su acceso público sin necesidad de instalación local. Esta plataforma facilita la ejecución de aplicaciones basadas en Gradio en entornos controlados y reproducibles, alineándose con enfoques abiertos de desarrollo de inteligencia artificial (Wolf et al, 2020).

Resultados

Los resultados obtenidos en este estudio se estructuran en torno a tres ejes analíticos principales:

1 el desempeño cuantitativo del modelo entrenado sobre el conjunto de prueba,

2 la interpretación detallada de la matriz de confusión y de las métricas por clase,

3 la validación funcional del sistema mediante su implementación como aplicación web interactiva.

Cada uno de estos ejes permite comprender, desde diferentes perspectivas, el alcance y las limitaciones del modelo propuesto, articulando consideraciones técnicas, clínicas y pedagógicas. A partir de esta estructura, la presente sección se organiza en cinco grandes apartados: rendimiento global, métricas por clase, análisis de errores, discusión sobre desempeño diagnóstico y evaluación del funcionamiento práctico en el entorno web.

Desempeño global del modelo

El modelo final, basado en ResNet18 y entrenado con técnicas de transferencia de aprendizaje sobre el *dataset* de Kermany et al (2018), alcanzó un nivel de desempeño satisfactorio en el conjunto de prueba. La ejecución del ciclo de inferencia sobre las 624 imágenes del *test set* produjo los siguientes resultados generales:

- *Accuracy*: 0,8942
- Macro F1-*score*: 0,8827
- *Weighted* F1-*score*: 0,8919

Estos valores reflejan un rendimiento sólido para una arquitectura de baja profundidad (18 capas) y un proceso de entrenamiento deliberadamente sencillo, orientado a la reproducibilidad y aplicabilidad educativa, más que a la búsqueda de máximos rendimientos como ocurre en desafíos competitivos o en entornos clínicos de alta especialización.

La *accuracy* cercana al 90% se encuentra dentro del rango esperado para modelos residuales entrenados en *dataset*s radiológicos acotados. En el estudio original de He et al (2016), las arquitecturas ResNet demostraron gran capacidad de generalización en tareas visuales diversas; sin embargo, el rendimiento en imágenes médicas depende fuertemente de la disponibilidad de datos y de la variabilidad clínica re-

presentada. En este sentido, los resultados son consistentes con investigaciones previas basadas en CNN aplicadas a neumonía, como las de Rajpurkar et al (2017), quienes obtuvieron mejoras significativas cuando trabajaron con arquitecturas mucho más profundas y con millones de parámetros adicionales.

Si bien este proyecto no persigue competir con modelos de frontera, los niveles de precisión alcanzados indican que la estrategia de *fine-tuning* sobre una arquitectura liviana constituye un camino válido para desarrollar entornos de enseñanza sobre IA en diagnóstico por imágenes. La relativa simplicidad del modelo permite también una comprensión más accesible para estudiantes de ciencias de la salud, quienes pueden explorar cómo las capas convolucionales identifican patrones radiográficos fundamentales.

Análisis de métricas por clase

Además de la evaluación agregada en términos de *accuracy*, resulta necesario analizar el desempeño del modelo en cada una de las clases. La clasificación binaria entre imágenes *NORMAL* y *PNEUMONIA* presenta desafíos particulares. La clase *PNEUMONIA* suele contener patrones heterogéneos (consolidaciones focales, opacidades difusas, intersticiopatías, etc.), mientras que la clase *NORMAL* puede incluir variaciones anatómicas que no constituyen patología, pero sí podrían inducir falsos positivos. A continuación, se presentan los resultados obtenidos, primeramente, en Tabla 1, se muestra el Reporte de Clasificación:

Tabla 1. Reporte de clasificación.

	Precisión	*Recall*	F1-*Score*	Soporte
NORMAL	0,9330	0,7735	0,8458	234
PNEUMONIA	0,8767	0,9667	0,9195	390
Accuracy (precisión)			0,8942	624
Promedio Macro (*macro avg*)	0,9049	0,8701	0,8827	624
Promedio ponderado (*Weighted avg*)	0,8978	0,8942	0,8919	624

Fuente: elaboración propia.

Clase NORMAL

- **Precisión:** 0,933
- **Recall:** 0,7735
- **F1-**score**:** 0,8458
- **Soporte:** 234 imágenes

Clase PNEUMONIA

- **Precisión:** 0,8767
- **Recall:** 0,9667
- **F1-**score**:** 0,9195
- **Soporte:** 390 imágenes

La primera observación es que el modelo es particularmente fuerte en la detección de casos de neumonía: su *recall* del 96,67% implica que el sistema identifica correctamente casi todos los casos patológicos. *Recall* o recuperación es también conocida como sensibilidad o tasa de verdaderos positivos, es una métrica de rendimiento en el aprendizaje automático que mide la capacidad de un modelo para identificar todas las instancias relevantes dentro de un conjunto de datos. Este aspecto resulta clínicamente relevante, dado que el objetivo en aplicaciones diagnósticas es minimizar los falsos negativos. En otras palabras, el modelo rara vez deja pasar inadvertido un caso de neumonía, atributo muy valioso en contextos clínicos donde una omisión podría retrasar tratamientos necesarios.

Sin embargo, este comportamiento favorable hacia la clase patológica viene acompañado de una caída en el *recall* para la clase NORMAL. Con un valor del 77,35%, significa que una proporción de radiografías sin patología se clasifican erróneamente como casos de neumonía. Este comportamiento es común en modelos entrenados sobre *dataset*s con leves desbalances o con alta variabilidad en un rango limitado de parámetros clínicos. Por otro lado, es positivo que emita falsos positivos, debido a que es preferible que un diagnóstico sobre estime la patología y luego la confirme. En cambio, un falso negativo puede ser devastador en la salud de un paciente.

Desde la perspectiva formativa, esta asimetría ofrece un excelente material pedagógico para discutir los sesgos y compensaciones inherentes a cualquier sistema automatizado. Permite también introducir conceptos cruciales en evaluación diagnóstica: la diferencia entre sensibilidad y especificidad, el impacto de los falsos positivos en la carga asistencial y el papel de la IA como herramienta complementaria, no sustitutiva, del juicio clínico profesional.

Matriz de confusión: análisis profundo de los patrones de error

La matriz de confusión ofrece una visión más detallada del comportamiento del modelo. Sus datos se expresan en la Tabla 2.

Tabla 2. Datos principales de la matriz de confusión del modelo.

	Predicción NORMAL	**Predicción** PNEUMONIA
Real NORMAL	181	53
Real PNEUMONIA	13	377

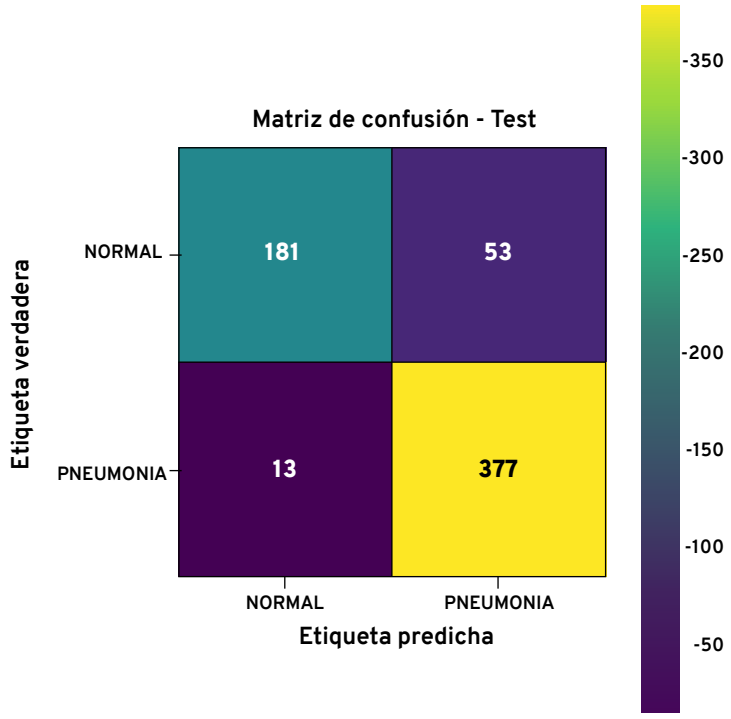
Fuente: elaboración propia.

Este desglose permite profundizar en los aciertos y errores del modelo:

- Verdaderos negativos (TN): 181
- Falsos positivos (FP): 53
- Falsos negativos (FN): 13
- Verdaderos positivos (VP): 377

Dicho desglose se presenta en la matriz de confusión que se expresa en la Figura 1.

Figura 1. Matriz de confusión del Modelo.



Fuente: elaboración propia.

El modelo presenta un patrón característico de muchos clasificadores en imágenes médicas: una marcada tendencia hacia la identificación correcta de la patología (VP elevados) y un número relativamente bajo de falsos negativos, pero al costo de un incremento de falsos positivos.

Esto puede interpretarse desde dos perspectivas:

Perspectiva clínica

Desde un punto de vista médico, es preferible un modelo que tienda al exceso de precaución (falsos positivos) antes que uno que omita casos reales (falsos negativos). La neumonía es un cuadro potencialmente grave, por tanto, un algoritmo que detecte de más puede ser aceptable si se utiliza como apoyo al diagnóstico.

Perspectiva técnica

Los falsos positivos pueden derivarse de:

- **Artefactos radiográficos:** sombras, subexposición o sobreexposición en bordes torácicos prominentes.
- **Patrones normales atípicos:** variaciones anatómicas benignas que no se encuentran bien representadas en el *dataset*.
- **Aprendizaje accidental de artefactos:** como descargas del equipo, estructuras externas o imperfecciones en el *dataset*.

Este análisis constituye un excelente insumo para discutir explicabilidad de modelos. La introducción futura de técnicas como Grad-CAM (Selvaraju et al, 2017) permitirá identificar si la red está focalizando en regiones anatómicas relevantes o si está influida por artefactos no clínicos.

Interpretación desde la literatura: desempeño comparado

Los resultados obtenidos deben analizarse a la luz de la literatura existente sobre redes neuronales aplicadas a neumonía. El estudio CheXNet (Rajpurkar et al, 2017), basado en DenseNet121, reportó niveles de F1 superiores, aunque con un conjunto de datos hospitalario mucho más extenso y diverso (112.000 radiografías). Del mismo modo, Wang et al (2017) y Irvin et al (2019) trabajaron con bases de datos masivas cuyo volumen habilita generalización más robusta.

En comparación, un modelo entrenado exclusivamente con ResNet18 y un *dataset* más pequeño no está diseñado para alcanzar los máximos *benchmarks de última generación* sino para ofrecer un equilibrio entre rendimiento, accesibilidad y explicabilidad pedagógica.

Dentro de ese enfoque formativo, la precisión (*accuracy*) del 89% resulta altamente significativa, en especial considerando que:

- la arquitectura es liviana;
- el dataset no supera las 6.000 imágenes;
- el entrenamiento fue breve;
- el objetivo es educativo y demostrativo.

Por ello, el desempeño obtenido se considera adecuado y consistente con los fines del proyecto.

Análisis pedagógico del desempeño

Uno de los aspectos más relevantes del presente estudio es su aplicabilidad en contextos educativos, particularmente en carreras de salud como Bioimágenes o Instrumentación Quirúrgica que son en las que se desempeña el plantel docente que participa del presente artículo. El modelo no solo permite mostrar resultados numéricos, sino que habilita discusiones profundas en torno a:

diferenciación entre sensibilidad y especificidad;

- impacto de los falsos negativos en contextos clínicos;
- interpretación crítica de matrices de confusión;
- comprensión de sesgos inherentes del *dataset*;
- principios de *overfitting* y *underfitting;*
- relevancia de la calidad de las imágenes.

El hecho de que el modelo marque más falsos positivos es una oportunidad para reflexionar sobre la necesidad de supervisión humana en cualquier herramienta de IA en salud. Permite desmontar la idea de automatización infalible y promover una alfabetización digital responsable, donde docentes y estudiantes puedan comprender las fortalezas y debilidades de las herramientas que utilizan.

Evaluación funcional mediante la aplicación web

La última parte de los resultados involucra la validación del modelo en un entorno operativo real: la aplicación web desarrollada con Gradio y desplegada en HuggingFace Spaces.

Durante la fase de pruebas, la aplicación demostró:

1. Correcta carga de imágenes de RX de tórax.
2. Procesamiento en tiempo real con preprocesamiento coherente al entrenamiento.
3. Predicciones consistentes con el modelo original exportado.

4. Visualización clara de la clase predicha, probabilidad asociada y precisión (*accuracy*) global.

Este componente constituye evidencia empírica de que el modelo no solo funciona en condiciones controladas (como los notebooks de Kaggle), sino que puede operar en un entorno realista donde usuarios externos cargan imágenes nuevas.

La interfaz web permite experimentar con imágenes reales sin necesidad de instalar software local, lo que facilita actividades docentes, demostraciones y estudios exploratorios en aulas de salud e informática.

A la herramienta e interfaz web se accede mediante un enlace público generado por HuggingFace Spaces.

https://huggingface.co/spaces/Lguer/Neumonia-Rx-UNPAZ

Implicancias para la investigación futura

El desempeño observado invita a explorar líneas futuras tales como:

- integración de Grad-CAM para explicabilidad visual;
- ampliación del dataset con imágenes (clasificadas por sexo, adultas o pediátricas y/o situadas, es decir, provenientes de hospitales locales de los partidos aledaños);
- entrenamiento con técnicas avanzadas como freeze-unfreeze escalonado o cyclical learning rates;
- comparación con arquitecturas más profundas (ResNet50, DenseNet121);
- exploración con modelos vision transformer (ViT);
- validación cruzada sobre múltiples datasets.

Estas líneas permitirán robustecer los resultados, aumentar la generalización y profundizar las posibilidades pedagógicas del sistema, ya que permitirán discutir y comparar resultados entre dichos modelos.

Discusión

Los resultados obtenidos permiten analizar el desempeño del modelo desde una perspectiva técnica, clínica y pedagógica, considerando tanto su alineación con el estado del arte como las condiciones metodológicas bajo las cuales fue desarrollado.

En términos de rendimiento, el modelo basado en ResNet18 alcanzó una *accuracy* cercana al 90%, con un *recall* superior al 96% para la clase *PNEUMONIA*. Este comportamiento resulta consistente con la literatura que destaca la capacidad de las redes convolucionales para identificar patrones radiográficos asociados a patologías pulmonares, incluso en contextos de variabilidad clínica (Rajpurkar et al, 2017; Wang et al, 2017; Irvin et al, 2019). No obstante, la elección de una arquitectura de menor

profundidad implica una limitación en la capacidad de generalización en comparación con modelos más complejos como DenseNet121 o ResNet50, particularmente cuando se trabaja con *dataset*s de tamaño moderado.

Desde el punto de vista técnico, la utilización de ResNet18 permite evidenciar un equilibrio entre desempeño y eficiencia computacional. Las redes residuales introducidas por He et al (2016) han demostrado facilitar el entrenamiento estable de modelos profundos mediante el uso de conexiones de salto. Sin embargo, la detección de patrones radiográficos sutiles puede requerir arquitecturas más profundas o estrategias de ajuste fino más avanzadas. En este estudio, la priorización de la reproducibilidad y la accesibilidad condiciona el rendimiento máximo alcanzable, pero ofrece un marco metodológico transparente y replicable.

En relación con la relevancia clínica, el elevado *recall* para la clase *PNEUMONIA* indica una alta capacidad del modelo para detectar casos patológicos, reduciendo la probabilidad de falsos negativos. Este comportamiento es consistente con la literatura clínica, que señala la importancia de priorizar la sensibilidad en sistemas de apoyo al diagnóstico (Topol, 2019). No obstante, la disminución del *recall* en la clase *NORMAL* evidencia la presencia de falsos positivos, lo que puede derivar en una sobreestimación de la patología. Si bien este comportamiento es aceptable en contextos de triage automatizado, plantea limitaciones en escenarios donde la sobrecarga asistencial es un factor crítico.

Estas diferencias entre clases también reflejan la influencia del *dataset* en el comportamiento del modelo. El conjunto de datos de Kermany et al (2018), si bien adecuado para fines formativos, presenta un desbalance a favor de la clase patológica y una variabilidad limitada en imágenes normales. Este sesgo puede inducir al modelo a interpretar patrones anatómicos normales como indicios de enfermedad. Asimismo, el predominio de imágenes pediátricas restringe la generalización a poblaciones adultas, lo que debe ser considerado al interpretar los resultados y su potencial aplicación en contextos clínicos reales.

Desde una perspectiva pedagógica, el desarrollo de un modelo reproducible utilizando herramientas gratuitas como PyTorch, Kaggle, Gradio y HuggingFace Spaces constituye un aporte significativo para la educación superior. La posibilidad de replicar un flujo completo de aprendizaje profundo –desde el entrenamiento hasta el despliegue– facilita la incorporación de la inteligencia artificial en carreras de ciencias de la salud y tecnología, especialmente en contextos con recursos limitados. Este enfoque permite a los estudiantes analizar métricas, interpretar matrices de confusión y reflexionar sobre los sesgos y limitaciones de los sistemas automatizados, promoviendo una alfabetización digital crítica.

Asimismo, el comportamiento del modelo ofrece oportunidades para discutir conceptos fundamentales en evaluación diagnóstica, tales como la relación entre sensibilidad y especificidad, el impacto de los falsos positivos y la necesidad de supervisión humana en el uso de herramientas de IA. En este sentido, el modelo no debe interpretarse como un sustituto del juicio clínico, sino como un recurso complementario que requiere validación contextual y análisis crítico.

Las limitaciones metodológicas condicionan la interpretación de los resultados. El tamaño y la composición del *dataset* restringen la capacidad de generalización del modelo, mientras que la ausencia de validación en conjuntos de datos externos limita la evaluación de su robustez en escenarios clínicos diversos. Asimismo, la falta de técnicas de explicabilidad, como Grad-CAM (Selvaraju et al, 2017), impide analizar si las decisiones del modelo se basan en regiones anatómicas relevantes o en artefactos no clínicos. Estas limitaciones señalan la necesidad de avanzar hacia modelos más robustos, validaciones multicéntricas e incorporación de mecanismos de interpretabilidad.

Los resultados evidencian que es posible desarrollar un modelo funcional, reproducible y pedagógicamente valioso utilizando herramientas accesibles, aunque su aplicación en entornos clínicos requiere precaución, validación adicional y mejoras metodológicas.

Líneas futuras

El presente estudio abre diversas líneas de desarrollo orientadas a fortalecer tanto la robustez técnica del modelo como su valor educativo y su potencial aplicación en contextos clínicos.

Una de las principales direcciones consiste en extender la propuesta hacia el análisis automatizado de otras modalidades de imágenes médicas, como mamografías, donde la inteligencia artificial ha mostrado avances significativos en la detección temprana de patologías. Este enfoque resulta especialmente relevante si se incorpora una perspectiva de género, considerando la incidencia del cáncer de mama y la necesidad de desarrollar herramientas que contemplen la diversidad de contextos sociodemográficos y condiciones clínicas. La construcción de *dataset*s representativos y el diseño de modelos sensibles a estas variaciones constituyen un desafío central para evitar la reproducción de sesgos en el diagnóstico.

Desde el punto de vista técnico, futuras investigaciones deberán explorar arquitecturas de mayor profundidad y capacidad representacional, tales como DenseNet121, ResNet50 o modelos basados en *Vision Transformers* (ViT), que han demostrado un mejor desempeño en tareas de clasificación radiológica compleja. Asimismo, resulta prioritario incorporar estrategias explícitas de mitigación del desbalance de clases, como funciones de pérdida ponderadas, muestreo balanceado o técnicas como *Focal Loss*, con el objetivo de mejorar el rendimiento en la clase *NORMAL* sin comprometer la sensibilidad diagnóstica.

Otra línea de avance fundamental radica en la incorporación de técnicas de explicabilidad, como Grad-CAM, que permitan visualizar las regiones de la imagen utilizadas por el modelo en su proceso de decisión. Este tipo de herramientas no solo contribuye a la validación técnica del sistema, sino que también resulta clave para su apropiación pedagógica y su eventual integración en entornos clínicos.

Metodológicamente, se plantea la necesidad de ampliar la validación del modelo mediante el uso de *dataset*s externos y multicéntricos, que incluyan mayor diversidad anatómica, etaria y técnica. Esta ampliación permitiría evaluar la capacidad de generalización del sistema en escenarios más cercanos a la práctica clínica real.

Desde una perspectiva educativa, se propone avanzar hacia el desarrollo de entornos integrados de aprendizaje en diagnóstico por imágenes, donde los modelos de inteligencia artificial puedan ser utilizados de manera progresiva y reflexiva. La articulación entre herramientas tecnológicas accesibles y propuestas pedagógicas situadas permitirá fortalecer la alfabetización digital crítica en estudiantes de ciencias de la salud, promoviendo una comprensión profunda de las potencialidades y limitaciones de la inteligencia artificial en contextos profesionales.

Conclusiones

El estudio presentado demuestra que la combinación de modelos de aprendizaje profundo preentrenados y un *pipeline* de entrenamiento cuidadosamente diseñado permite desarrollar un sistema capaz de clasificar radiografías de tórax en dos categorías fundamentales para la práctica clínica: normales y compatibles con neumonía. La utilización de ResNet-18, junto con un proceso de *fine-tuning* optimizado y validado experimentalmente, evidenció que incluso arquitecturas relativamente livianas pueden alcanzar desempeños robustos cuando se las entrena con un *dataset* adecuadamente curado y se controlan las condiciones del preprocesamiento, la segmentación y la evaluación. La precisión obtenida en el conjunto de prueba (≈89%) no solo se alinea con estudios previos, sino que confirma que es posible construir herramientas educativas y demostrativas sin requerir una infraestructura de cómputo avanzada.

Asimismo, el desarrollo y despliegue de una aplicación web funcional, basada en Gradio y posteriormente integrable en Hugging Face Spaces, demuestra la viabilidad técnica de acercar modelos de IA a entornos de formación profesional en carreras vinculadas al diagnóstico por imágenes. Esta etapa es particularmente relevante, ya que permite que estudiantes y docentes interactúen directamente con el modelo, comprendan sus fortalezas y limitaciones y desarrollen criterios críticos respecto del uso real de estas tecnologías en el ámbito clínico. El resultado es un puente significativo entre la investigación computacional y la práctica pedagógica, contribuyendo a democratizar el acceso a herramientas de inteligencia artificial aplicadas a la salud.

El trabajo confirma también la importancia de diseñar procesos de validación explícitos y transparentes, incorporando métricas como matriz de confusión, precisión por clase, *recall* y F1-*score*. Estas métricas ofrecen una comprensión más profunda que la mera exactitud global, permitiendo identificar fortalezas diferenciales del modelo –como su elevado *recall* para neumonía– y áreas donde su rendimiento es perfectible –particularmente en la clasificación de radiografías normales. De este modo, se favorece una interpretación adecuada del modelo, evitando lecturas ingenuas o sobreconfianza injustificada.

Del análisis integral del desarrollo, se desprende una conclusión central: las herramientas basadas en IA para el análisis de imágenes médicas constituyen un recurso complementario, no sustitutivo, en los procesos de diagnóstico. La precisión del modelo, aun siendo elevada, no alcanza los niveles de especificidad y sensibilidad necesarios para un uso clínico sin supervisión. Sin embar-

go, su valor como tecnología educativa, como soporte experimental en investigación formativa y como prototipo demostrativo para entender el funcionamiento de los sistemas de visión por computadora es indiscutible.

Este proyecto aporta a un debate más amplio en torno a la incorporación de IA en salud, donde convergen cuestiones técnicas, éticas, pedagógicas y de soberanía tecnológica. El trabajo pone de manifiesto que la apropiación crítica de estas tecnologías requiere no solo modelos precisos, sino también desarrollos que integren perspectivas de equidad, explicabilidad y accesibilidad. En ese sentido, la propuesta constituye un punto de partida prometedor para futuras investigaciones orientadas a patologías más complejas, modelos más robustos y entornos multicéntricos reales, consolidando un camino de articulación entre inteligencia artificial, educación superior y salud pública.

Referencias bibliográficas

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. https://www.tensorflow.org/

Gradio Team. (2024). *Gradio: Build machine learning apps in Python* (Version 5.x). https://www.gradio.app/

He, K., Zhang, X., Ren, S. y Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., … Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 590-597. https://doi.org/10.1609/aaai.v33i01.3301590

Kermany, D., Zhang, K. y Goldbaum, M. (2018). *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification* (Version 2) [Dataset]. Mendeley Data. https://doi.org/10.17632/rscbjbr9sj.2

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M. Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (Vol. 32). https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., … Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*. https://arxiv.org/abs/1711.05225

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. y Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. En *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618-626). https://doi.org/10.1109/ICCV.2017.74

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). *ChestX-ray14: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases*. National Institutes of Health Clinical Center.Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6