

# Explicabilidad y Detección de Shortcut Learning en Clasificación de Neumonía por Rayos X

Proyecto Final - Ethics and Explainable AI

Diciembre 2025

## Resumen

Este trabajo investiga la aplicabilidad de técnicas de Inteligencia Artificial Explicable (XAI) en la detección de neumonía mediante imágenes de rayos X torácicos. Se entrena un modelo baseline basado en ResNet-18 preentrenado y se compara con un modelo intencionalmente afectado por *shortcut learning*. Mediante técnicas como Grad-CAM y Activation Maximization, se demuestran diferencias en las características aprendidas por ambos, y cómo éstas no son evidentes mediante métricas tradicionales (como *accuracy* en un cierto conjunto de datos). Los resultados subrayan la importancia crítica de XAI en aplicaciones médicas para garantizar que los modelos aprendan patrones clínicamente relevantes.

## 1. Introducción

La aplicación de deep learning en diagnóstico médico ha mostrado resultados prometedores, pero plantea desafíos éticos y de confiabilidad [4]. Los modelos pueden aprender correlaciones espurias (*shortcut learning*) [1] en lugar de patrones clínicamente relevantes, comprometiendo su generalización y seguridad.

Este proyecto tiene tres objetivos principales: (1) desarrollar un modelo de clasificación de neumonía interpretable, (2) demostrar experimentalmente el fenómeno de shortcut learning mediante corrupción controlada de datos, y (3) validar técnicas XAI para detectar estos comportamientos no deseados.

## 2. Datos y Metodología

### 2.1. Dataset

Se utiliza el dataset Chest X-Ray Pneumonia de Kaggle [3], que contiene 5.856 radiografías torácicas en escala de grises con dos clases: NORMAL (27 %) y PNEUMONIA (73 %). La distribución presenta un desbalance significativo de 2.71:1 a favor de la clase PNEUMONIA.

Los datos se dividen en 85 % entrenamiento (4,434 imágenes), 15 % validación (782 imágenes) y un conjunto de test fijo (624 imágenes). Las imágenes presentan dimensiones variables (384-2916 píxeles de ancho), requiriendo redimensionamiento a  $224 \times 224$  píxeles para compatibilidad con ResNet-18.

### 2.2. Arquitectura del Modelo

Se emplea ResNet-18 preentrenado en ImageNet [2], modificando la capa final para clasificación binaria. El preentrenamiento proporciona características visuales robustas que facilitan el transfer learning. Durante el entrenamiento se aplica:

- Data augmentation: flip horizontal ( $p=0.5$ ), rotación ( $\pm 10^\circ$ ), jitter de brillo/contraste

- Optimizador: AdamW con learning rate inicial de  $10^{-4}$  y decaimiento de peso (*weight decay*) de  $10^{-4}$
- Función de pérdida: CrossEntropyLoss con pesos de clase balanceados
- Early stopping con paciencia de 10 épocas

### 2.3. Generación de Shortcut Learning

Para estudiar el fenómeno de forma controlada, se crea un segundo modelo entrenado desde cero (sin preentrenamiento <sup>1</sup>) sobre un dataset corrupto. Se añaden marcadores circulares blancos en la esquina superior izquierda únicamente a imágenes de la clase PNEUMONIA durante entrenamiento. Este artefacto artificial actúa como un *shortcut* espurio correlacionado perfectamente con la clase objetivo.

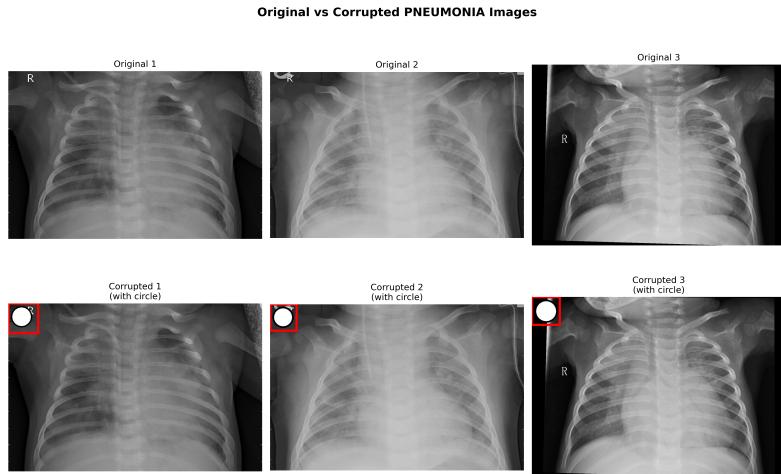


Figura 1: Ejemplos de imágenes originales vs. corruptas con marcador artificial

### 2.4. Entrenamiento y Evaluación

Ambos modelos (baseline y shortcut) se entrena, respectivamente, durante 20 y 15 épocas hasta convergencia. De cara a diagnosticar un posible sobreajuste, se monitorean las curvas de pérdida y accuracy en entrenamiento y validación (Figura 2a). Adicionalmente, se evalúa el rendimiento en datos limpios y corruptos para cuantificar el impacto del shortcut learning. Como podemos observar en la Figura 2b, el modelo baseline alcanza un accuracy del 81.57 % en datos limpios, mientras que el modelo shortcut logra un 98.88 % en datos corruptos pero sólo un 46.47 % en datos limpios, evidenciando su dependencia del marcador artificial.

## 3. Técnicas XAI Aplicadas

De cara a interpretar y comparar ambos modelos, hemos seleccionado dos técnicas de XAI, una local/híbrida (Grad-CAM) y otra global (Activation Maximization):

<sup>1</sup>En pruebas iniciales, se intentó hacer con preentrenamiento, pero el modelo no aprendía los atajos deseados (probablemente debido a que ya había aprendido características robustas que impedían la dependencia de atajos espurios durante el preentrenamiento)

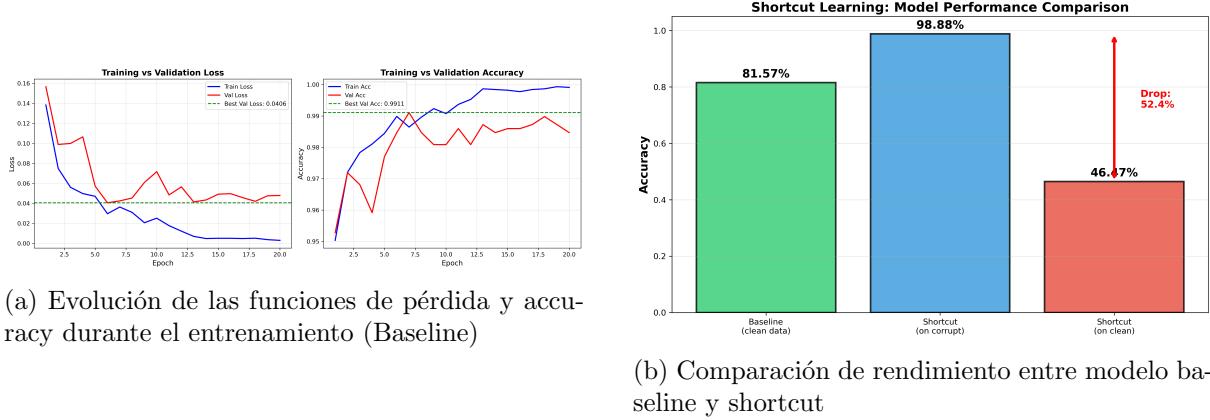


Figura 2: Análisis de sobreajuste y rendimiento entre modelos

### 3.1. Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM visualiza las regiones de una imagen concreta que más contribuyen a la decisión del modelo mediante el cálculo de gradientes de la clase objetivo respecto a los mapas de activación de la última capa convolucional. Los resultados muestran diferencias claras entre ambos modelos (Figura ??).

El modelo baseline enfoca correctamente en regiones pulmonares centrales, mientras que el modelo corrupto concentra su atención exclusivamente en la esquina superior izquierda donde se ubica el marcador artificial.

### 3.2. Activation Maximization

Esta técnica genera imágenes sintéticas que maximizan la activación de cada clase mediante ascenso de gradiente. Para el modelo baseline, los prototipos de la clase NORMAL muestran patrones de costillas y estructuras anatómicas reconocibles, mientras que los prototipos de PNEUMONIA presentan texturas complejas de bajo nivel menos interpretables.

En contraste, el modelo shortcut genera prototipos sin estructuras anatómicas claras. Los scores de confianza son significativamente menores (2.5 para NORMAL, -2.6 para PNEUMONIA) comparados con el baseline (13 para NORMAL, 40 para PNEUMONIA). El score negativo para PNEUMONIA indica la incapacidad del modelo para generar ejemplos de esta clase sin el marcador artificial.

### 3.3. Análisis de Atención en Múltiples Ejemplos

El análisis sistemático de Grad-CAM sobre múltiples muestras revela patrones consistentes:

- **Modelo Baseline:** Atención distribuida en áreas pulmonares bilaterales, identificando opacidades y consolidaciones coherentes con presentación clínica de neumonía
- **Modelo Shortcut:** Atención invariablemente concentrada en la región del marcador artificial, independientemente del contenido médico de la imagen

## 4. Sanity Checks y Actionable Insights

### 4.1. Validación mediante Sanity Checks

Se implementan dos sanity checks fundamentales para validar la fidelidad de las explicaciones XAI:

**1. Randomization Test:** Se reemplazan progresivamente los pesos del modelo por valores aleatorios, comenzando desde las capas superiores. Las explicaciones del modelo baseline pierden coherencia gradualmente conforme aumenta la aleatorización, confirmando que Grad-CAM refleja genuinamente el proceso de decisión aprendido.

**2. Baseline Comparison:** Se comparan las explicaciones con un modelo con pesos completamente aleatorios. El modelo entrenado muestra mapas de atención estructurados y consistentes, mientras que el modelo aleatorio produce mapas de ruido sin patrones discernibles.

## 4.2. Métricas de Rendimiento

El contraste entre ambos modelos es evidente en el desempeño sobre datos limpios vs. corruptos:

Modelo	Acc. Datos Limpios	Acc. Datos Corruptos
Baseline	87.5 %	-
Shortcut	46.5 %	98.9 %

Cuadro 1: Comparación de rendimiento entre modelos

La precisión del modelo shortcut en datos limpios (46.5 %) es inferior incluso a un clasificador aleatorio (50 %), confirmando que el modelo no aprendió características relevantes de neumonía sino únicamente la correlación con el marcador artificial.

## 4.3. Actionable Insights para Aplicaciones Médicas

Este estudio genera recomendaciones prácticas:

- 1. Validación obligatoria con XAI:** Todo modelo médico debe evaluarse con técnicas de explicabilidad antes del despliegue clínico
- 2. Inspección de datos:** Los datasets médicos deben auditarse para detectar correlaciones espurias (marcadores institucionales, artefactos de adquisición, sesgos demográficos)
- 3. Transfer learning con precaución:** El preentrenamiento en ImageNet aporta robustez contra shortcuts, pero no elimina completamente el riesgo
- 4. Evaluación en distribuciones múltiples:** Los modelos deben validarse en datos de diferentes instituciones y equipos radiológicos

## 5. Limitaciones y Discusión

### 5.1. Limitaciones del Estudio

**Interpretabilidad humana vs. fidelidad:** Las visualizaciones Grad-CAM son plausibles para humanos pero pueden no capturar completamente el razonamiento del modelo. La diferencia entre *faithfulness* y *plausibility* es una limitación inherente a técnicas de explicabilidad en visión por computadora.

**Simplificación experimental:** El shortcut introducido (marcador circular) es deliberadamente obvio. Los shortcuts reales en datasets médicos suelen ser más sutiles (diferencias en contraste institucional, textos en imágenes, patrones de compresión).

**Tamaño del dataset:** Con 5,856 imágenes, el dataset es relativamente pequeño para estándares de deep learning. Estudios clínicos requerirían validación en decenas de miles de casos.

## 5.2. Implicaciones Éticas

El fenómeno de shortcut learning en IA médica plantea serias preocupaciones éticas. Un modelo que clasifica basándose en artefactos en lugar de patología real podría:

- Fallar catastróficamente al ser desplegado en instituciones con diferentes equipos/protocolos
- Generar falsa confianza en profesionales médicos al mostrar alta precisión en validación
- Perpetuar sesgos si los shortcuts correlacionan con demografías específicas

La obligatoriedad de XAI en sistemas médicos de alto riesgo no es solo una mejor práctica técnica, sino un imperativo ético para garantizar seguridad del paciente.

## 6. Conclusiones

Este proyecto demuestra experimentalmente la efectividad de técnicas XAI para detectar shortcut learning en clasificación de imágenes médicas. Los resultados principales son:

1. El modelo baseline basado en transfer learning aprende características anatómicas relevantes, evidenciado por Grad-CAM que enfoca en regiones pulmonares y Activation Maximization que genera prototipos con estructuras reconocibles
2. El modelo shortcut, entrenado desde cero con datos corruptos, depende exclusivamente del marcador artificial, alcanzando 98.9% de precisión en datos corruptos pero solo 46.5% en datos limpios
3. Los sanity checks validan que las explicaciones XAI reflejan fielmente el razonamiento aprendido por los modelos
4. El preentrenamiento en ImageNet proporciona robustez contra shortcuts simples, aunque no es una solución universal

En conclusión, XAI no es opcional en aplicaciones médicas de IA: es una necesidad técnica y ética para garantizar que los modelos aprendan patrones clínicamente relevantes en lugar de correlaciones espurias. La combinación de múltiples técnicas de explicabilidad (Grad-CAM, Activation Maximization, sanity checks) proporciona una visión más completa y confiable del comportamiento del modelo.

## A. Ilustraciones Adicionales

Si bien no incluimos todas las figuras en el cuerpo principal del informe (en pos de la brevedad), proponemos una serie de ilustraciones complementarias que pueden facilitar la comprensión de los conceptos tratados:

## Referencias

- [1] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

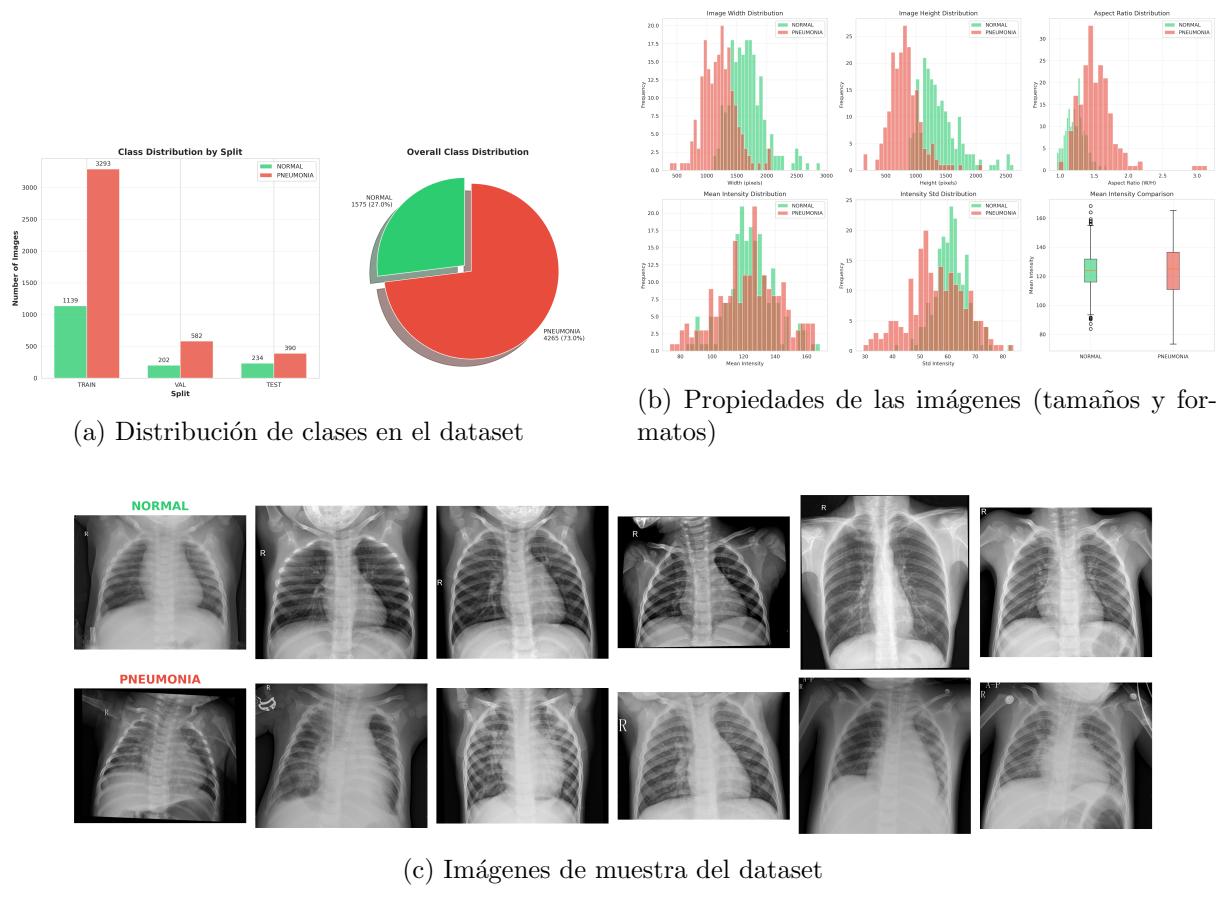


Figura 3: Información visual adicional sobre el dataset utilizado

- [3] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification. Kaggle, 2018. Accessed: December 2025.
- [4] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.