

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

Proyecto Final - Ética y Explicabilidad de la IA
Francisco Javier Ríos

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. Todos los códigos y datos están disponibles en https://github.com/Javirios03/final_project_xai.

1. Introducción

La aplicación de deep learning en diagnóstico médico ha mostrado resultados prometedores, pero plantea desafíos éticos y de confiabilidad [1]. Los modelos pueden aprender correlaciones espurias (*shortcut learning*) [2] 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 224×224 píxeles para compatibilidad con ResNet-18.

2.2. Arquitectura del Modelo

Se emplea ResNet-18 preentrenado en ImageNet [4], 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 3 épocas

2.3. Generación de Shortcut Learning

Para estudiar el fenómeno de forma controlada, se crea un segundo modelo entrenado desde cero (sin preentrenamiento ¹) 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.

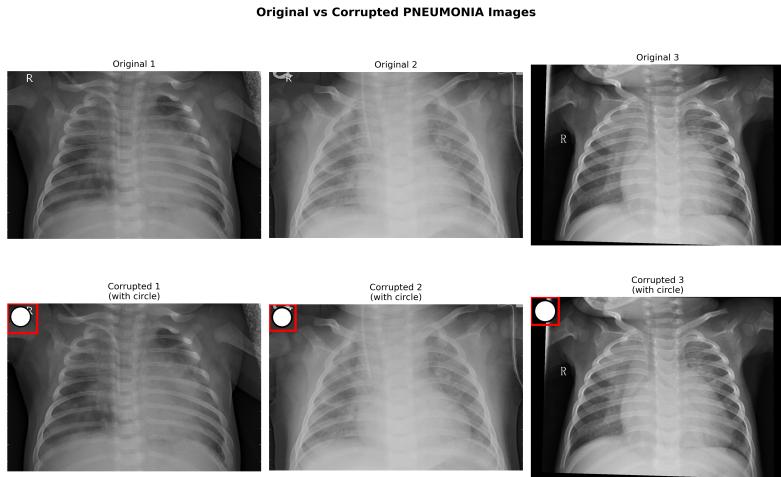


Figura 1: Ejemplos de imágenes originales vs. corruptas con marcador artificial

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.

¹En 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)

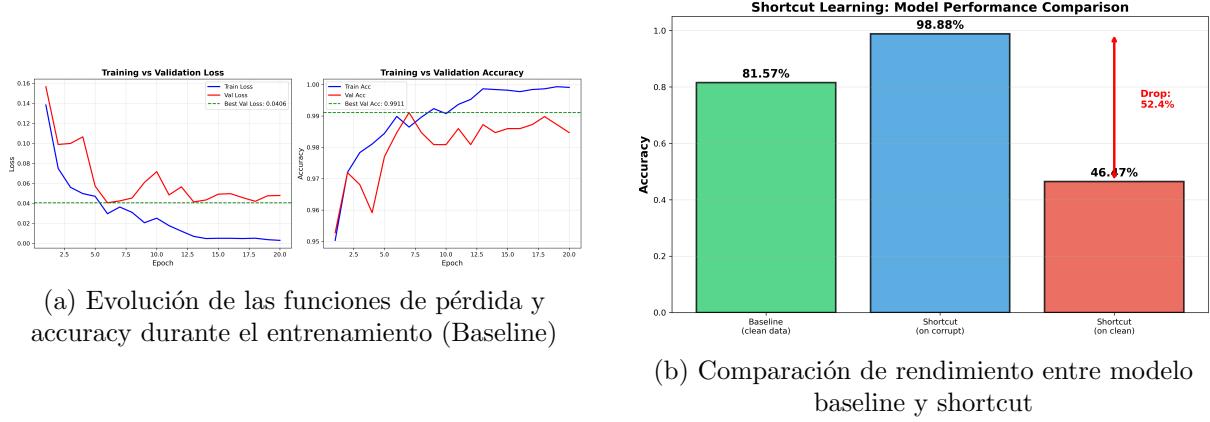


Figura 2: Análisis de sobreajuste y rendimiento entre modelos

Algo a destacar es que, si bien el modelo baseline alcanza un accuracy que no es sobresaliente (81.57 %), sus métricas de recall (99.23 %) y AUC (0.9487) son elevadas (B.1), lo que indica que el modelo es capaz de identificar casi todos los casos positivos de neumonía, aunque a costa de una mayor tasa de falsos positivos. Esto es crucial en aplicaciones médicas donde la sensibilidad es prioritaria (el coste derivado de falsos negativos es mucho mayor que el de falsos positivos).

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):

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

Grad-CAM [5] 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. De esta forma, se generan mapas de calor superpuestos a la imagen original que indican las áreas de mayor atención del modelo.

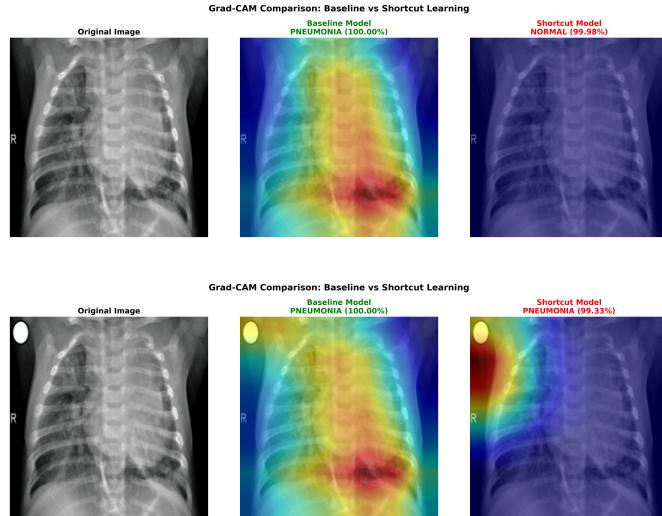


Figura 3: Comparación de mapas Grad-CAM para un caso de PNEUMONIA: (arriba) imagen limpia, (abajo) imagen corrupta con marcador artificial

El modelo baseline enfoca correctamente en el área pulmonar, específicamente el lóbulo inferior derecho, donde se observan opacidades características de neumonía. Dicha atención no

cambia radicalmente al introducir el marcador artificial. En contraste, en el caso de la imagen corrupta, el modelo shortcut concentra toda su atención en el marcador circular y, en la imagen limpia, no muestra atención significativa en ninguna región relevante.

En adición a este análisis cualitativo, realizamos un estudio cuantitativo de Grad-CAM sobre múltiples ejemplos, obteniendo un mapa de calor promediado a lo largo de 100 imágenes de la clase PNEUMONIA:

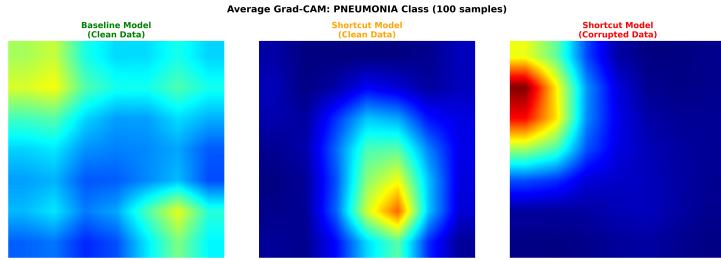


Figura 4: Mapa de calor promedio de Grad-CAM sobre 100 imágenes de PNEUMONIA

El modelo baseline muestra atención distribuida por todo el campo pulmonar, sin focos de activación extrema, lo cual es consistente con la variabilidad clínica de la neumonía. Sin embargo, el modelo shortcut presenta un comportamiento totalmente diferente. Es interesante analizar cómo cambia la atención al comparar imágenes limpias y corruptas: en imágenes limpias, parece centrarse en el cuadrante inferior derecho (lo cual parece indicar que, si bien el grueso de lo aprendido es el marcador, el modelo también intenta buscar alguna característica médica en ausencia del atajo). En imágenes corruptas, la atención se concentra exclusivamente en la esquina superior izquierda, donde se encuentra el marcador artificial, ignorando completamente el resto de la imagen.

Este análisis refuerza la conclusión de que el modelo shortcut no aprende características médicas relevantes (que el baseline sí capta), eliminando ruido de casos específicos. Sin embargo, debemos tener en cuenta las limitaciones inherentes a Grad-CAM: aunque es intuitivo para humanos, no garantiza una fidelidad absoluta al proceso de decisión del modelo [6]. Adicionalmente, obtener el mapa de calor promedio puede diluir detalles importantes sobre la heterogeneidad de casos individuales. Esto implica que no podemos distinguir, a falta de un análisis más profundo, si nuestros modelos miran consistentemente regiones relevantes o si alterna aleatoriamente entre distintas áreas o estrategias.

3.2. Activation Maximization

Si bien Grad-CAM ofrece una visión local del comportamiento del modelo, necesitamos responder a la pregunta: ¿qué características globales ha aprendido el modelo para representar cada clase? Para ello, utilizamos Activation Maximization [7, 8], una técnica que sintetiza imágenes que maximizan la activación de una clase específica en el modelo, produciendo un prototípo visual de lo que el modelo asocia con dicha clase.

En la figura superior, observamos 4 imágenes sintetizadas por clase y modelo. Analizándolas por separado:

- **Modelo Baseline:**

- *Clase NORMAL:* Las imágenes sintetizadas muestran estructuras que recuerdan costillas (líneas blancas horizontales y verticales) con una cierta simetría bilateral y apariencia anatómica (se parecen, si bien vagamente, a una radiografía torácica normal). Esto parece indicar que el modelo ha aprendido que la presencia de estas estructuras

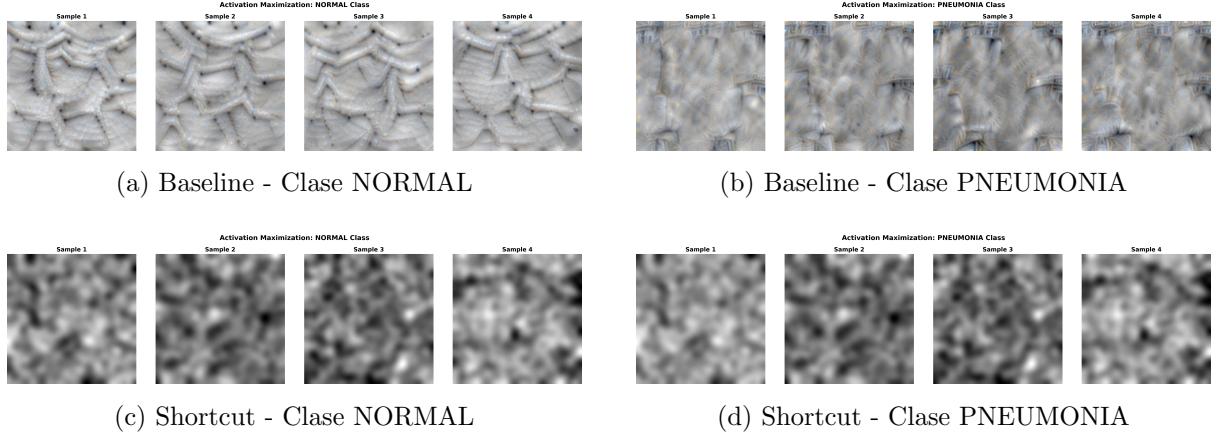


Figura 5: Imágenes sintetizadas mediante Activation Maximization para ambas clases y modelos

es indicativa de una radiografía normal (en la cual no existen opacidades ni anomalías visibles que oculten dichas estructuras).

- **Clase PNEUMONIA:** Cabe destacar que dichas imágenes son relativamente uniformes y una intensidad media-baja. Esto sugiere que el modelo se guía por texturas complejas de bajo nivel que no son fácilmente interpretables por el ojo humano, lo cual conecta con una de las mayores limitaciones al tratar de explicar modelos de visión por ordenador: esa diferencia tan marcada entre *faithfulness* y *plausibility* que estudiamos a lo largo del curso. Intentar simplificar el razonamiento del modelo a mecanismos antropocéntricos (formas reconocibles, objetos familiares) puede ser engañoso, ya que el modelo puede estar utilizando patrones estadísticos complejos que no son evidentes para humanos.
- **Modelo Shortcut:** Respecto al modelo shortcut, los prototipos generados para ambas clases son muy similares entre sí y presentan patrones que más bien parecen ruido aleatorio, sin estructuras reconocibles ni características anatómicas claras. Esto refuerza la conclusión de que el modelo no ha aprendido características médicas relevantes, sino que se basa en el marcador artificial.

Adicionalmente, es importante mencionar los *scores* obtenidos durante la optimización de Activation Maximization (ver Apéndice C). El modelo baseline alcanza scores significativamente más altos (40 para PNEUMONIA y 13 para NORMAL) en comparación con el modelo shortcut (2.5 para NORMAL y -2.5 para PNEUMONIA). Esto indica que el modelo baseline tiene una representación más robusta y diferenciada de las clases, mientras que el modelo shortcut no logra generar activaciones fuertes, reflejando su dependencia en un atajo espurio.

No sólo eso, sino que el hecho de que el score para la clase PNEUMONIA en el modelo shortcut sea negativo (-2.5) sugiere que el modelo tiene una dificultad extrema para generar imágenes de esa clase sin la presencia del marcador artificial. Analizando esto más en detalle, dado que el método de Activation Maximization aplica regularizaciones y desenfoques periódicamente, dicha optimización no es capaz de encontrar una solución tan específica que requiere:

1. La presencia del marcador artificial (un círculo perfecto)
2. Con una intensidad muy concreta (blanco puro)
3. En una localización concreta (esquina superior izquierda)

4. Sanity Checks y Actionable Insights

4.1. Validación mediante Sanity Checks

De cara a validar la fidelidad de las explicaciones obtenidas mediante Grad-CAM, hemos implementado un *sanity check* basado en la metodología propuesta por Adebayo et al. [6], consistente en randomizar los pesos del modelo² y observar cómo cambian las visualizaciones de Grad-CAM:

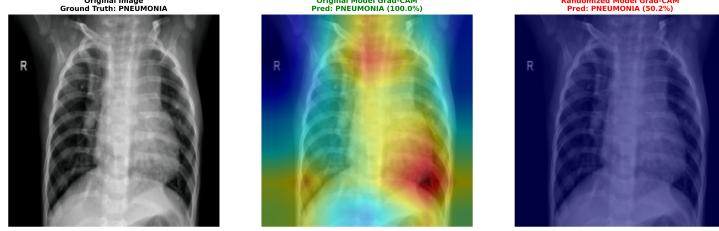


Figura 6: Sanity check de Grad-CAM mediante randomización de pesos

De este análisis, concluimos que las visualizaciones de Grad-CAM son efectivamente sensibles a los pesos del modelo. La confianza del modelo en su predicción colapsa a un 50,2% (en esencia, un clasificador aleatorio en este problema binario) tras la randomización, y las regiones de atención se vuelven muy débiles. Todo ello indica que las explicaciones obtenidas reflejan fielmente el razonamiento aprendido por el modelo, reforzando la validez de los análisis previos.

4.2. Aplicaciones prácticas para Aplicaciones Médicas

Los resultados detallados en las anteriores secciones subrayan la importancia crítica de incorporar técnicas XAI en el desarrollo y despliegue de modelos de IA en contextos médicos. Basándonos en estos hallazgos, proponemos las siguientes recomendaciones prácticas:

1. **Validación obligatoria con XAI:** El uso de técnicas como Grad-CAM y Activation Maximization deja de ser un extra útil para convertirse en un requisito indispensable en la validación de modelos médicos
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

²En este caso, nos limitamos a aplicarlo al modelo baseline, ya que el modelo shortcut no presenta mapas de atención coherentes desde el inicio

5. Limitaciones y Discusión

5.1. Limitaciones del Estudio

Si bien los resultados obtenidos derivan en conclusiones valiosas, es importante reconocer las limitaciones inherentes a nuestro proyecto:

5.1.1. Interpretación humana frente a fidelidad del modelo

Las visualizaciones de Grad-CAM y las imágenes sintetizadas mediante Activation Maximization, si bien visualmente intuitivas, presentan una limitación inherente: el privilegio de la plausibilidad frente a la fidelidad al razonamiento interno. Los mapas de atención muestran regiones de alta activación en *feature maps* de capas convolucionales, pero dicha activación es una correlación, no necesariamente una prueba causal del proceso de decisión del modelo.

Como se discute en [6], el uso indiscriminado de explicaciones visuales puede inducir a conclusiones erróneas, especialmente si los usuarios finales (médicos, reguladores) no están plenamente conscientes de las limitaciones de estas técnicas. Adicionalmente, existen algunas restricciones metodológicas relacionadas con las técnicas empleadas. Por ejemplo, el uso de la activación ReLU en Grad-CAM elimina activaciones negativas, lo cual puede ocultar regiones que el modelo podría usar como evidencia en contra de una clase.

Respecto a Activation Maximization, la optimización de imágenes sintetizadas depende fuertemente de las regularizaciones aplicadas (desenfoque, penalizaciones de norma). Asimismo, dichos prototipos muestran texturas de bajo nivel que no son fácilmente interpretables por humanos, lo cual limita su utilidad práctica en entornos clínicos (mostrando la otra cara de la moneda del trade-off entre plausibility y faithfulness).

5.1.2. Simplificación experimental

El uso de un marcador artificial para inducir shortcut learning, si bien efectivo para demostrar el fenómeno, es una simplificación que puede no capturar la complejidad de shortcuts reales en datos médicos. En escenarios clínicos, dichos shortcuts suelen no ser tan obvios y pueden involucrar múltiples factores sutiles (diferencias en protocolos de adquisición, demografía del paciente, artefactos específicos de equipos). Esta simplificación facilita la demostración del fenómeno, pero puede limitar la generalización de los hallazgos a situaciones del mundo real.

5.1.3. Tamaño y diversidad del dataset

Por otro lado, el dataset utilizado, aunque adecuado para propósitos experimentales, presenta limitaciones en cuanto a tamaño y diversidad. Un total de 5.856 imágenes puede no ser representativo de la variabilidad clínica completa de la neumonía, y la distribución desbalanceada de clases podría influir en el aprendizaje del modelo (en este caso, induciendo un sesgo intrínseco hacia la clase mayoritaria, PNEUMONIA). Estudios futuros deberían considerar datasets más grandes y diversos para validar la robustez de las conclusiones.

5.1.4. Ausencia de Validación Clínica

Finalmente, es crucial destacar que este estudio carece de validación clínica directa. Si bien las técnicas XAI proporcionan insights valiosos sobre el comportamiento del modelo, la interpretación y utilidad de estas explicaciones debe ir acompañada de la experiencia de profesionales médicos (a la cual debemos remitir, en última instancia, para justificar la relevancia de nuestros hallazgos). Por lo tanto, es nuestra opinión que toda conclusión obtenida mediante modelos de IA debe ser supeditada a la revisión y validación por parte de expertos clínicos.

5.2. Implicaciones Éticas

El fenómeno estudiado a lo largo de este proyecto tiene profundas implicaciones éticas en el contexto de la IA médica. La dependencia de shortcuts espurios puede conducir a consecuencias potencialmente desastrosas:

- Fallar catastróficamente al ser desplegado en instituciones con diferentes equipos/protocolos
- Generar falsa confianza en profesionales médicos al mostrar alta precisión en validación
- Comprometer la seguridad del paciente al basar decisiones en correlaciones no clínicas

De igual modo, la ciega confianza en explicaciones XAI sin una comprensión crítica de sus limitaciones puede inducir a errores de interpretación. La falta de esto último puede radicar en una falsa sensación de seguridad, llevando a decisiones clínicas erróneas basadas en explicaciones engañosas.

6. Conclusiones

Nuestro proyecto demuestra experimentalmente la efectividad de técnicas XAI para detectar shortcut learning en clasificación de imágenes médicas. Sintetizamos algunos hallazgos clave:

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:

A.1. Visualización del Dataset

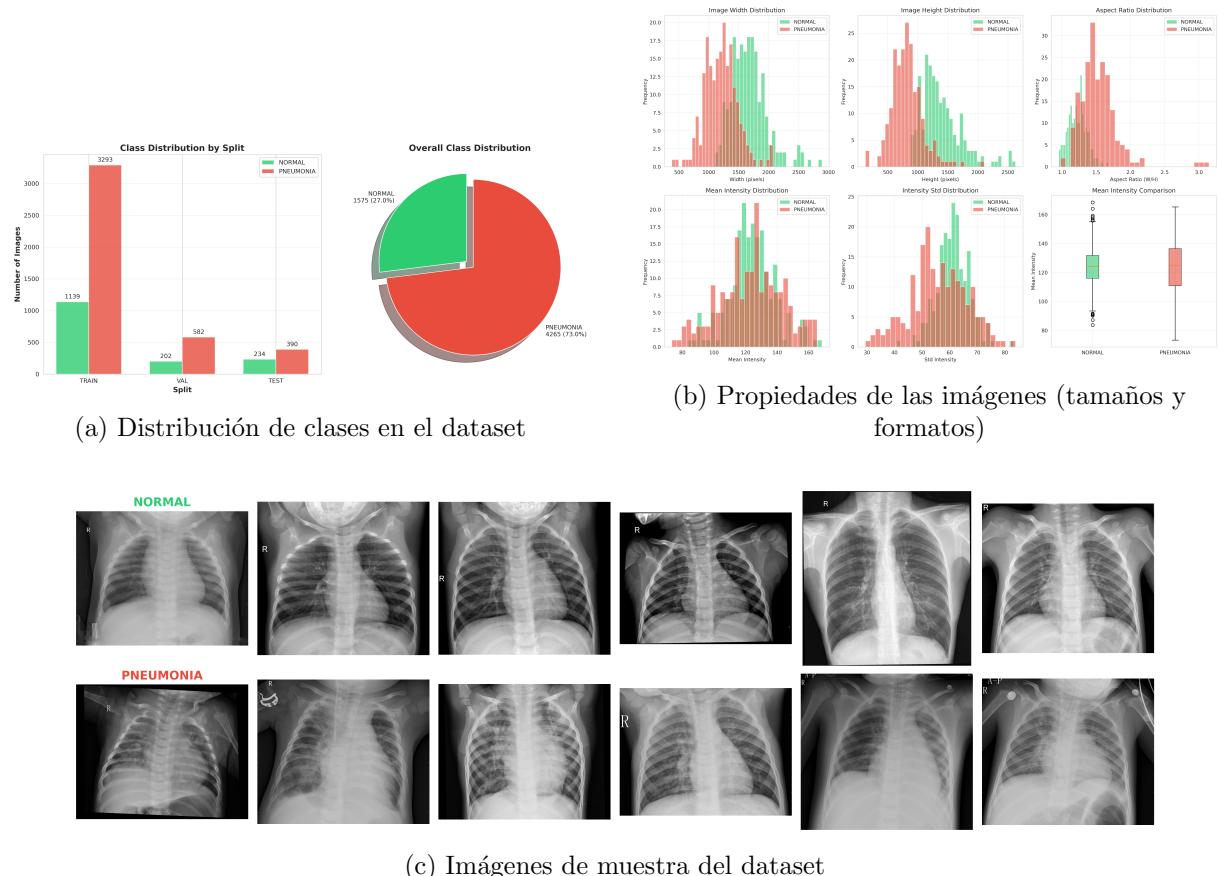


Figura 7: Información visual adicional sobre el dataset utilizado

A.2. Visualizaciones de Grad-CAM Complementarias

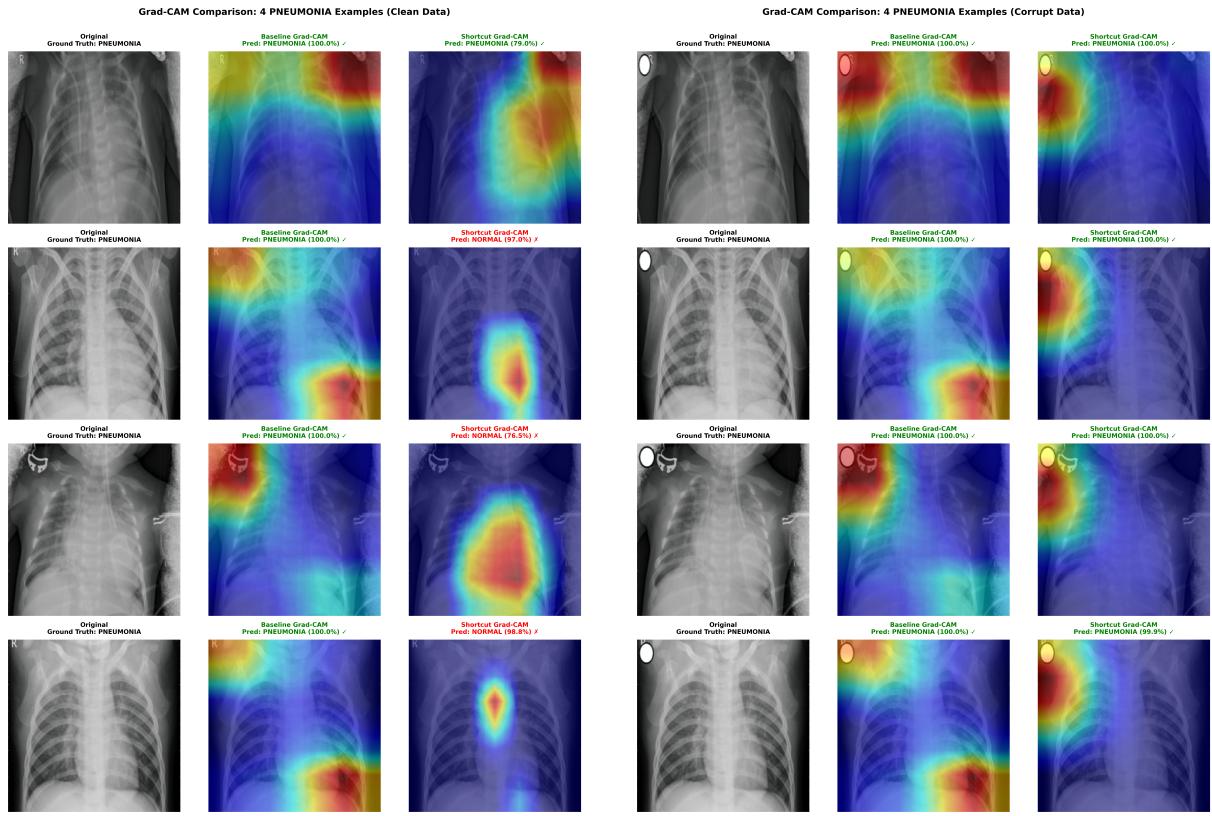


Figura 8: Visualizaciones de Grad-CAM complementarias en casos de neumonía: (izquierda) imágenes limpias, (derecha) imágenes corruptas

A.3. Distribución de la atención Grad-CAM según regiones pulmonares

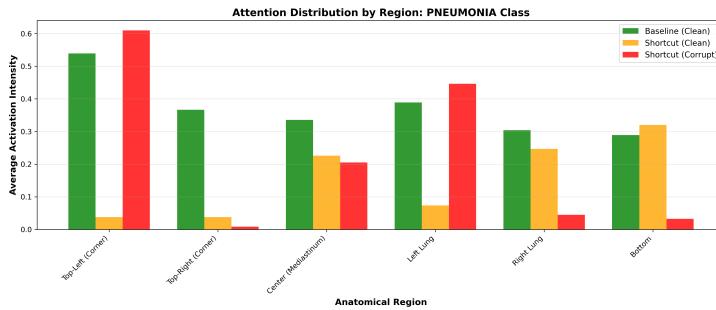


Figura 9: Distribución cuantitativa de la atención Grad-CAM en diferentes regiones pulmonares

B. Métricas Detalladas de Rendimiento

B.1. Métricas del Modelo Baseline

Métrica	Valor
Accuracy	81.57 %
Precision	77.56 %
Recall	99.23 %
F1-Score	87.06 %
AUC	0.9487

Cuadro 1: Métricas detalladas del modelo baseline en el conjunto de test limpio

B.2. Matriz de Confusión

	Pred: NORMAL	Pred: PNEUMONIA
True: NORMAL	186 (79.5 %)	48 (20.5 %)
True: PNEUMONIA	3 (0.8 %)	387 (99.2 %)

Cuadro 2: Matriz de confusión del modelo baseline en test set

C. Scores de Activation Maximization

Los scores obtenidos (medidos en logits) para las imágenes sintetizadas mediante Activation Maximization son los siguientes:

Modelo	Clase	Score
Baseline	NORMAL	13
Baseline	PNEUMONIA	40
Shortcut	NORMAL	2.5
Shortcut	PNEUMONIA	-2.5

Cuadro 3: Scores de Activation Maximization para cada modelo y clase

Referencias

- [1] 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.
- [2] 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.
- [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] 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.
- [5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.
- [6] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 9505–9515, 2018.
- [7] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 2009.
- [8] Hongbo Zhu and Angelo Cangelosi. Representation understanding via activation maximization, 2025.