

# Inteligencia Artificial Explicable aplicada a un Modelo de CNN para el Diagnóstico de Neumonía Infantil a partir de Radiografías Torácicas

David Álvarez Rojas  
*david.alvarez@usm.cl*

Sebastián Jara Cifuentes  
*sebastian.jara@alumnos.usm.cl*

**Abstract**— En el presente documento se introduce el problema del diagnóstico de neumonía infantil en contextos clínicos y las razones por las cuales es importante mirar con urgencia este problema. Luego se propone la realización de una investigación para mejorar el diagnóstico de la neumonía infantil mediante el uso de modelos de Machine Learning (ML), más específicamente Redes Neuronales Convolucionales (CNN). En este mismo contexto, se presentan objetivos de la investigación, el conjunto de datos a usar para entrenar el modelo propuesto, y métodos de Inteligencia Artificial Explicable (XAI) para mejorar la interpretabilidad del modelo de caja negra; que posteriormente se aplican para explicar las predicción hechas por un modelo CNN entrenado.

**Palabras claves**—XAI, CNN, Neumonía Infantil, Machine Learning, Modelos de Caja Negra, LIME, ShapValues, Smooth-Grad, RISE.

## I. INTRODUCCIÓN

La neumonía es una enfermedad grave que causa muchas muertes, especialmente infantiles, en países en desarrollo, superando las causadas por el VIH/SIDA; la malaria; y el sarampión juntas [2].

Diagnosticar y tratar correctamente la neumonía es crucial para un infante. Un diagnóstico rápido de la enfermedad puede ser la diferencia entre la vida y la muerte del menor. El “gold standard” actual es utilizar radiografías torácicas de los pulmones para la detección de la neumonía. Sin embargo, en lugares con recursos limitados, a menudo no se dispone de un acceso rápido y de alta calidad a expertos radiólogos que interpreten correctamente estas imágenes. Esto es un problema, ya que es necesario distinguir entre neumonía bacteriana y viral para un tratamiento adecuado.

Por otro lado, el uso de redes neuronales, y en particular las redes neuronales convolucionales (CNN), han demostrado ser una herramienta muy valiosa en el ámbito del diagnóstico clínico. Estos modelos de machine learning (ML) pueden analizar una amplia gama de datos clínicos, como imágenes de radiografías, resonancias magnéticas, etc. identificando patrones sutiles que a menudo escapan al ojo humano experto y que pueden permitir diagnosticar de manera rápida y precisa. Sin embargo, a pesar de su eficacia, estas redes neuronales a menudo se consideran cajas negras en el contexto clínico. Esto significa que, si bien pueden brindar diagnósticos precisos, es difícil entender y explicar cómo llegaron a esa decisión. La falta de transparencia en estos modelos de caja negra plantea preocupaciones éticas significativas en la comunidad médica,

ya que los médicos y los pacientes confían en la justificación y la interpretación de los resultados de diagnóstico para tomar decisiones informadas sobre la atención médica y el tratamiento.

De lo anterior, resulta necesario en el contexto médico explicar como un modelo de caja negra usado para diagnóstico o tratamiento, toma una decisión. Por lo tanto, es pertinente y crucial abordar este desafío implementando técnicas de inteligencia artificial explicables (XAI) que permitan comprender y verificar las decisiones tomadas por los modelos de aprendizaje profundo.

## II. DESCRIPCIÓN DEL PROBLEMA

Con el propósito de lograr un diagnóstico certero, transparente y comprensivo de la neumonía infantil, la presente investigación propondrá un modelo basado en CNN para apoyar el diagnóstico de neumonía a partir de radiografías torácicas del paciente. Posteriormente, con el modelo entrenado, se implementarán técnicas de inteligencia artificial explicables (XAI) que permitirán una comprensión más clara y transparente de cómo el modelo toma decisiones y realiza la clasificación.

Además de buscar lograr una alta precisión en las tareas de clasificación de imágenes, resulta necesario que las decisiones del modelo sean interpretables, lo que es crucial en aplicaciones clínicas y de diagnóstico de enfermedades, donde se requiere una justificación y una alta confiabilidad en las predicciones realizadas por el modelo. En este contexto, resulta fundamental buscar respuestas a las siguientes preguntas: ¿Cómo pueden las técnicas de XAI mejorar la comprensión de las decisiones de un modelo de CNN en el diagnóstico de neumonía a partir de radiografías de tórax? ¿Qué impacto tendría la interpretabilidad de las decisiones del modelo en la confiabilidad de los resultados de diagnóstico en aplicaciones clínicas? ¿Cuáles son las técnicas de XAI más adecuadas para integrar en un modelo de CNN en el contexto del diagnóstico de neumonía?

### A. Hipótesis de Investigación

La hipótesis de investigación de este proyecto es que la aplicación de técnicas de XAI permiten una comprensión más clara de las decisiones que toma un modelo de CNN en el diagnóstico de neumonía, permitiendo entregar una

justificación interpretable a las predicciones del modelo, y así mejorando la confiabilidad del diagnóstico.

#### B. Objetivo General

El objetivo general de esta investigación es implementar técnicas de XAI a un modelo de CNN de diagnóstico de neumonía en infantes usando radiografías torácicas, esto con el fin de mejorar interpretabilidad del modelo.

#### C. Objetivos Específicos

- Construir o re-usar un modelo de CNN para que diagnostique la presencia de neumonía a partir de radiografías torácicas.
- Entrenar el modelo usando un conjunto de datos de radiografías torácicas de infantes.
- Identificar y seleccionar técnicas de XAI adecuadas para el modelo.
- Integrar las técnicas de XAI post-hoc en el modelo de manera efectiva.
- Generar visualizaciones y explicaciones de las predicciones del modelo.

#### D. Factibilidad

El presente estudio tiene una gran factibilidad debido a que existe una gran cantidad de datos para entrenar el modelo CNN para el diagnóstico de neumonía infantil. También es factible mejorar la interpretabilidad del modelo de clasificación, ya que al ser imágenes radiográficas, se pueden generar interpretaciones visuales, que por lo general pueden ser más interpretables por los usuarios.

#### E. Apoyo experto

Las dificultades de los autores sobre los conocimientos médicos relacionados al diagnóstico de enfermedades, como la neumonía infantil, serán suplidas a través del contacto con expertos del área de Biomédica de la Universidad de Valparaíso.

### III. TRABAJOS RELACIONADOS

Tras revisar un Review Paper del estado del arte en el uso de XAI [9] se identificaron (y seleccionaron) los siguientes trabajos relacionados que abordan la aplicación de la inteligencia artificial en el diagnóstico de enfermedades a partir de radiografías torácicas junto con el uso de métodos de XAI para mejorar la explicabilidad e interpretabilidad de las predicciones de las distintas arquitecturas:

- **Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach [7]:** Este trabajo utiliza una CNN para extraer las características de radiografías torácicas para poder realizar predicciones sobre la presencia de cáncer de pulmón a partir de imágenes de Rayos X Pulmonares, para luego emplear **Grad-CAM** para explicar las regiones relevantes de las imágenes que contribuyen a la predicción.
- **COV-ELM classifier: An extreme learning machine based identification of COVID-19 using chest X-ray**

#### images [8]:

En este trabajo se busca identificar COVID-19, mediante ML, a partir imágenes de radiografías torácicas. Para realizar esto se emplea un clasificador basado en Extreme Learning Machine [10], el cual es un algoritmo de aprendizaje rápido en redes neuronales que inicializa capas ocultas de manera aleatoria y encuentra pesos eficientemente a través de una operación de inversión de matriz. Una vez realizada la predicción se utiliza **LIME** para visualizar y corroborar los resultados con hallazgos clínicos.

- **Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays [6]:** Se utiliza una arquitectura de Deep Learning para detectar distintas enfermedades pulmonares y COVID-19 a partir de radiografías torácicas, para luego hacer uso de **Grad-CAM** para resaltar las regiones de interés que apoyan la predicción del modelo, y de esta forma explicar las predicciones de manera visual.

### IV. CONFIGURACIÓN EXPERIMENTAL

#### A. Conjunto de Datos

Se empleará el conjunto de datos “Chest X-ray” [1]. En este conjunto se ha reunido y etiquetado un conjunto de 5.232 imágenes de radiografías de tórax de niños. De estas, 3.883 muestran casos de neumonía (2.538 de origen bacteriano y 1.345 de origen viral), mientras que 1.349 son radiografías de pulmones normales (Ver figura 1). Estas imágenes provienen de un grupo de 5.856 pacientes y serán usadas para entrenar y validar un modelo de redes neuronales convolucionales.



Fig. 1. La radiografía torácica de pulmones normales (izquierda) muestra pulmones claros sin anomalías. La neumonía bacteriana (centro) muestra una mancha en el lóbulo superior derecho ( indicado por flechas blancas), mientras que la neumonía viral (derecha) muestra un patrón difuso en ambos pulmones.

#### B. Software y Hardware utilizado

El lenguaje de programación usado ha sido Python. Entre las bibliotecas utilizadas, destacan numpy y pandas para la manipulación y análisis de datos, tensorflow para la construcción y entrenamiento de la CNN, y matplotlib para la visualización de datos. Además, se incluye sklearn para funciones de modelado y selección de datos, y la biblioteca especializada de LIME para la interpretación de los modelos de aprendizaje automático. Estas bibliotecas proporcionaron un entorno adecuado para implementar los aspectos propuestos en el presente proyecto.

Respecto al hardware, debido a la poca profundidad de la red propuesta, fue suficiente utilizar los recursos que ofrece Google

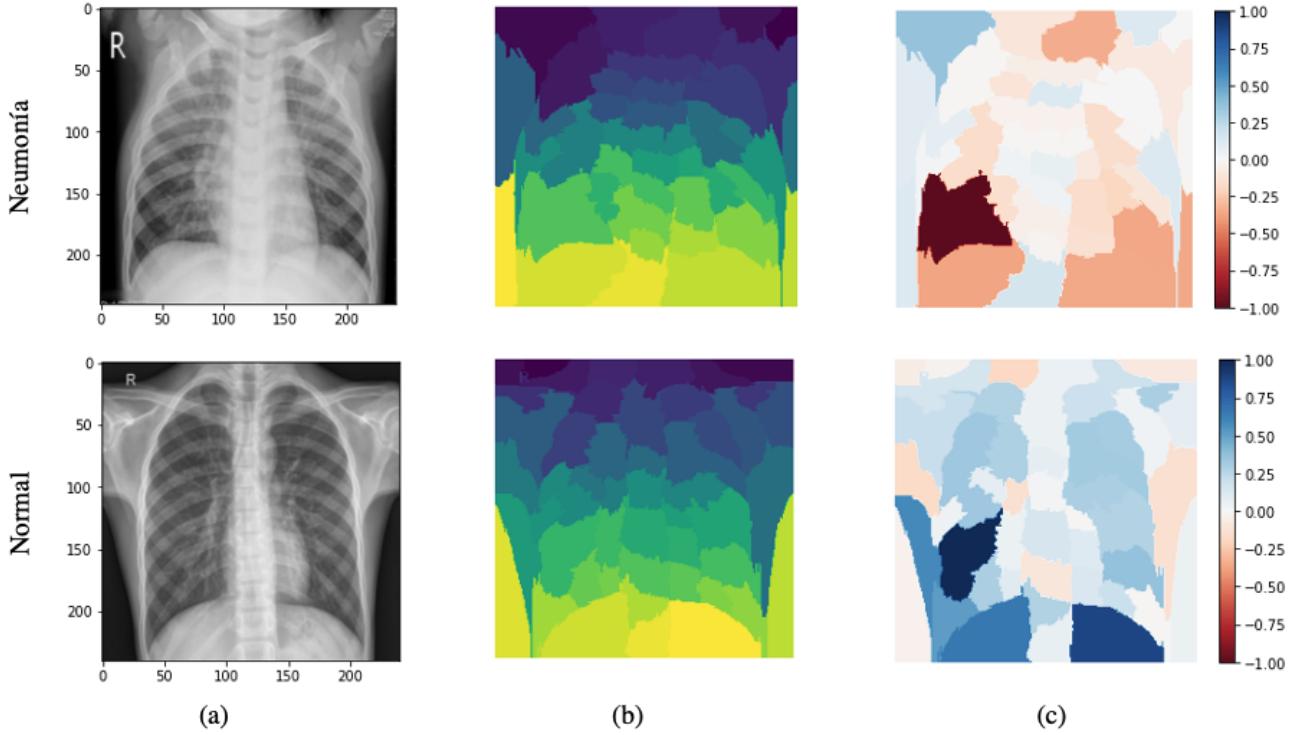


Fig. 2. Resultados aplicando el método LIME. En la columna (A) se presenta las radiografías correspondientes a un paciente con neumonía (arriba) y a un paciente sin neumonía (abajo). La columna (b) muestra los megapíxeles creados para la visualización de el mapa de prominencia mostrado en la columna (c).

Colab para el entrenamiento de la CNN. La plataforma Google Colab, con su acceso a GPUs, facilitó un eficiente entorno de entrenamiento sin la necesidad de recursos computacionales avanzados propios.

## V. METODOLOGÍA

En este proyecto, se empleará un modelo de redes neuronales Convolucionales de Clasificación de radiografías médicas torácicas para detectar y visualizar la presencia de neumonía en las radiografías torácicas presentes en el conjunto de datos [1]. El modelo que se propone está fuertemente inspirado en el modelo presentado en [3]. Observamos que es de particular importancia lograr altos niveles de rendimiento predictivo para la detección de neumonía, pues esta es la clase predominante disponibles en el conjunto de datos.

Una vez entrenado el modelo, y alcanzado una precisión sobre el 80% en el conjunto de validación, se aplicarán distintos métodos de explicabilidad sobre las predicciones del modelo CNN.

Finalmente, se comparan los métodos de explicabilidad en función de métricas automáticas y de señalamiento humano.

### A. Arquitectura de CNN

La red neuronal convolucional (CNN) diseñada sigue una arquitectura secuencial usando Keras, una API de alto nivel para modelos de aprendizaje automático.

Se usó Transfer Learning en la red, usando de modelo base *VGG16* [12] inicializada con los pesos de su entrenamiento

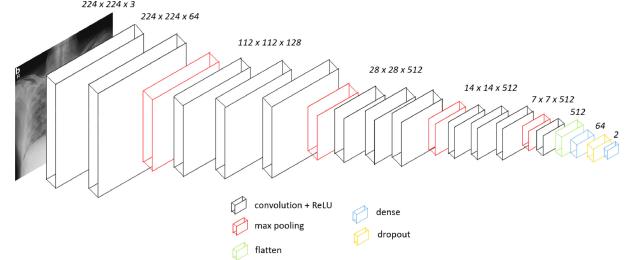


Fig. 3. Imagen sacada de [6]

en *Image-Net* [13]. Esta elección fue basada en el paper de Brunese y Mercaldo [6], donde ocupan esta red para solucionar el problema. La capa de entrada de la red recibe como input imágenes de 240 de alto, 240 de ancho con 3 canales (RGB). Estas imágenes pasan a través de cinco capas convoluciones con un tamaño de *kernel* de 3x3, *stride* 1, y con una *relu* como función de activación. Estas capas están seguidas por una capa de *MaxPooling* que reduce la dimensionalidad espacial y ayuda a reducir la carga computacional; los tres bloques convolucionales de la red son de 64, 128, 256, y los dos últimos de 512 filtros, respectivamente.

Una vez la imagen pasa por las capas convolucionales, la salida es aplanada a un vector unidimensional por una capa *Flatten* para poder pasar esta información extraída a la capas densas de 128 neuronas, seguido de una capa de *Dropout* para mejorar la generalización del modelo.

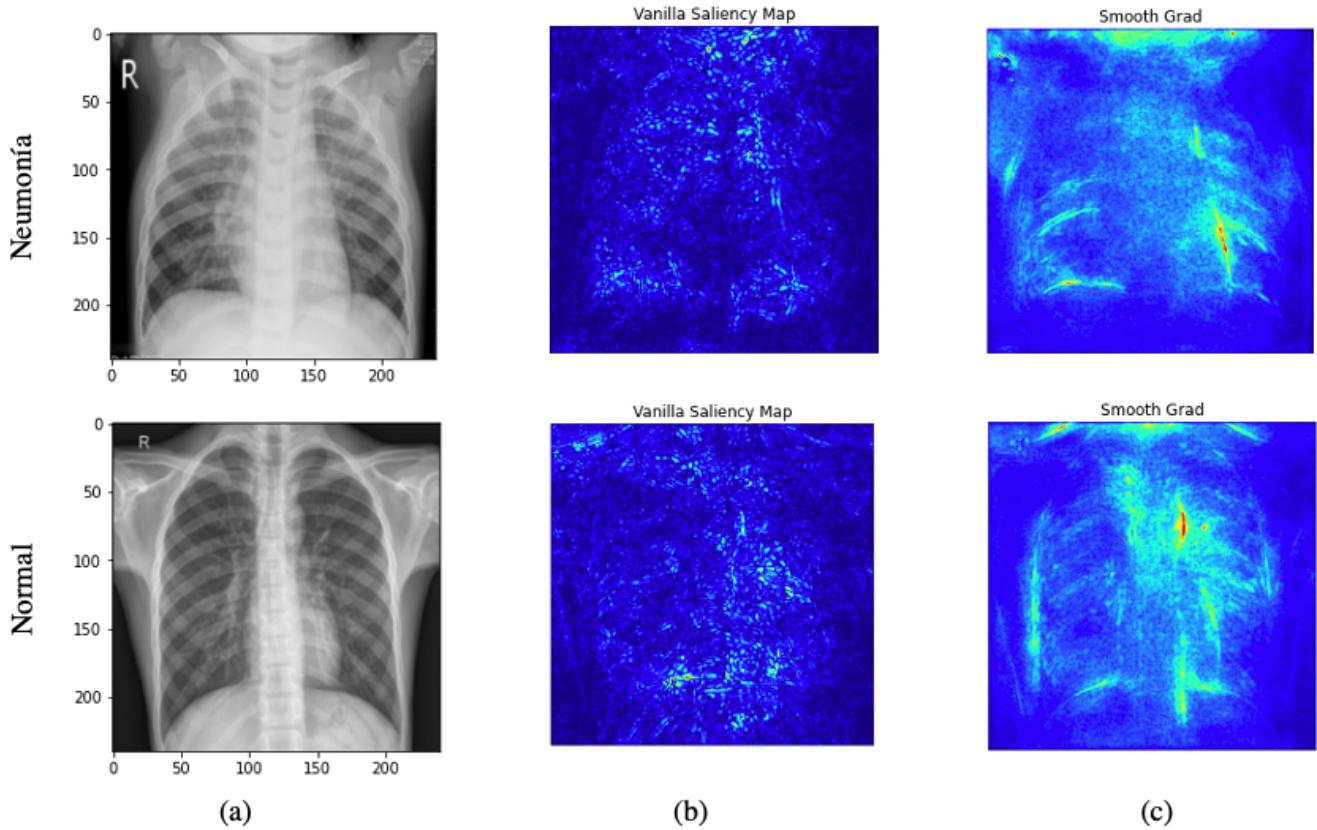


Fig. 4. Resultados aplicando el método Saliency Map y SmoothGrad. En la columna (a) se presenta las radiografías correspondientes a un paciente con neumonía (arriba) y a un paciente sin neumonía (abajo). La columna (b) y (c) representa las zonas de atención del modelo para realizar la predicción. En (b) se aplica el método Saliency Maps y en (c) SmoothGrad.

Finalmente, la capa de salida se compone de una capa Dense con 2 unidades y activación *softmax*, lo que permite a la red estar diseñada para resolver un problema de clasificación binaria (ver Figura 3).

#### B. Métodos de XAI

Para mejorar la interpretabilidad del modelo se utilizaron los siguiente métodos:

- **LIME** [4]: Es un método para explicar las predicciones de modelos de aprendizaje automático complejos, conocidos como “cajas negras”. Lo hace creando un modelo sustituto simple y interpretable, que aproxime al modelo original, para una instancia de datos específica (explicación local), para luego generar muestras de datos perturbados alrededor de la instancia de interés en el modelo simple, ponderadas por su proximidad.
- **Kernel SHAP** [4]: Es un método que calcula Shapley Values de forma aproximada para explicar la importancia de las características en un modelo de aprendizaje automático en una región local alrededor de un punto de datos de referencia. Los Shapley Values indican cuánto contribuye cada característica a una predicción dada, basándose en la Teoría de Juegos cooperativos. La explicación que

realiza KernelSHAP se hace asumiendo que el modelo es lineal, y no tiene en cuenta las interacciones entre las características. Este método proporciona una forma eficiente de comprender cómo las características afectan las predicciones del modelo en un contexto específico.

- **Saliency Maps** [4]: Son un método de explicación local que se basan en una expansión de la serie Taylor alrededor de una instancia de interés para aproximar la función de puntuación no lineal. Los Saliency Weights se calculan como las derivadas parciales de la función de puntuación con respecto a las características. Los Saliency Maps indican la importancia de cada característica para una predicción específica y son útiles para comprender la contribución de las características en una predicción local, pero no capturan interacciones entre características. Junto a los Vanilla Saliency Maps se ocupó una técnica llamada **SmoothGrad** [5], la cual es una técnica que busca mejorar la interpretabilidad de los Saliency Maps agregándole ruido a la imagen original, para luego promediar los Saliency Maps de las imágenes con distintos niveles de ruido.
- **GradCam** [14]: Es un método de explicabilidad para CNNs en tareas de clasificación de imágenes, siendo

una generalización de otro método llamado CAM. Para realizar la explicación, GradCAM elige la última capa convolucional para su análisis generando un mapa de localización discriminatorio mediante el cálculo de las derivadas parciales de las clases (antes de la capa output) con respecto a los mapas de activación de la capa convolucional, para luego realizar un *Global Average Pooling* de estas derivadas para obtener los **Pesos de Importancia** de cada neurona. Finalmente, se realiza una combinación ponderada de los mapas de activación, utilizando los pesos de importancia. Como GradCAM solo busca explicar los píxeles de la imagen que aportan a la predicción, le aplica una *ReLU* a esta combinación lineal para eliminar todas las “aportaciones negativas”.

- **Rise** [11]: El método RISE (Randomized Input Sampling for Explanations) es una técnica de explicabilidad Post-hoc, agnóstica, local y que no requiere acceso a las capas internas o pesos del modelo, lo que lo hace aplicable a una amplia gama de modelos de aprendizaje profundo. RISE estima la importancia de las regiones de la imagen del input para la predicción del modelo. A pesar de su simplicidad y generalidad, el método supera los enfoques de explicación existentes en términos de métricas causales automáticas y funciona de manera competitiva en términos de la métrica de señalización centrada en el ser humano, como se reporta en [11].

### C. Métricas Causales

Aún no hay consenso sobre cómo medir la explicabilidad de un modelo de aprendizaje automático [11]. Como resultado, la evaluación humana ha sido la forma predominante de evaluar la explicación, donde se presupone transparencia, confianza del usuario o la comprensión humana de las decisiones tomadas por el modelo. Los métodos de justificación existentes evalúan mapas de prominencia por su capacidad para localizar objetos. Sin embargo, la localización es simplemente un sustituto de la explicación humana y puede no capturar correctamente lo que causa que el modelo base tome una decisión, independientemente de si la decisión es correcta o incorrecta en lo que respecta a la tarea. Los autores de [11] argumentan que mantener a los humanos fuera del bucle de evaluación lo hace más justo y fiel a la propia visión del clasificador sobre el problema, en lugar de representar la visión de un humano. Tal métrica no solo es objetiva (libre de sesgo humano) por naturaleza, sino que también ahorra tiempo y recursos.

Siguiendo lo presentado en [11], se proponen dos métricas de evaluación automáticas: eliminación e inserción. La idea detrás de la métrica de eliminación es que la eliminación de la ‘causa’ obligará al modelo de clasificación a cambiar su decisión. Específicamente, esta métrica mide una disminución en la probabilidad de la clase predicha a medida que se eliminan más y más píxeles importantes, donde la importancia se obtiene del mapa de prominencia. Una caída abrupta y por lo tanto un menor área bajo la curva de probabilidad (como función de la fracción de píxeles eliminados) significa una buena explicación. La métrica de inserción, por otro lado, toma un

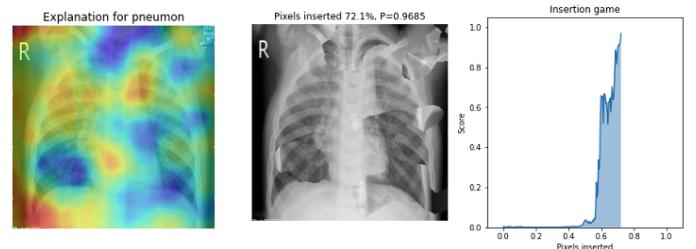


Fig. 5. Método de Inserción: Inicialmente se tiene una imagen que el modelo clasifica fuertemente como “normal”. En función de la importancia de los píxeles descrita en el mapa de prominencia (imagen de la izquierda), se comienzan a agregar píxeles de la imagen clasificada como un caso de neumonía (imagen del centro). El gráfico de la derecha muestra la evolución de la curva de probabilidad vs la proporción de píxeles insertados.

enfoque complementario. Mide el aumento en la probabilidad a medida que se introducen más y más píxeles, con un área bajo la curva de probabilidad más alto, indicativo de una mejor explicación.

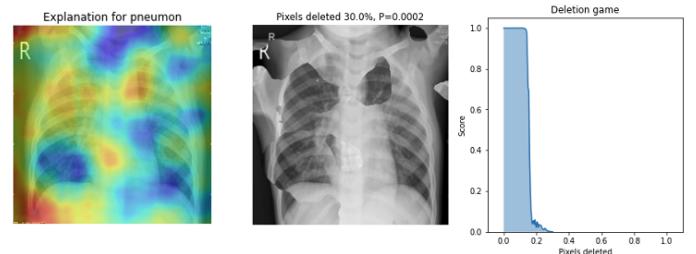


Fig. 6. Método de Borrado: Inicialmente se tiene una imagen clasificada como un caso de neumonía. En función de la importancia de los píxeles descrita en el mapa de prominencia (imagen de la izquierda), se comienzan a eliminar píxeles de la imagen inicial (imagen del centro). El gráfico de la derecha muestra la evolución de la curva de probabilidad vs la proporción de píxeles eliminados

Existen varias formas de eliminar píxeles de una imagen, por ejemplo, estableciendo los valores de los píxeles a cero o cualquier otro valor gris constante, desenfocando los píxeles o incluso recortando un cuadro delimitador ajustado. Estos enfoques fueron implementados sin éxito, ya que como el modelo propuesto predice la probabilidad sobre dos clases, una imagen inicial borrosa o una imagen completamente de un color, llevaba al modelo a decidir sobre una de estas dos clases, la mayoría de los casos el modelo predecía “neumonía” con probabilidad cercana a uno. Para resolver el problema anterior, se propone una manera de cálculo de esta métricas que sirve para las imágenes torácicas. Para la métrica de eliminación se toma una imagen de la clase neumonía, que haya sido correctamente predicha, para luego ir eliminando las zonas más prominentes que aportan a la decisión del modelo (definidas por el método a evaluar), siendo reemplazadas estas zonas por las zonas en la misma posición en una imagen predicha correctamente como normal. De forma contraria pero parecida, en el método de inserción se selecciona una imagen predicha correctamente como normal, para luego agregar los píxeles de las zonas más prominentes de una imagen predicha de manera correcta como neumonía. Esta forma de calcular las métricas de evaluación

funciona debido a la naturaleza de los datos utilizados y de imágenes de rayos x, no siendo generalizable con otros datasets. o inicial (para la métrica de inserción) una radiografía que el modelo clasifique fuertemente como “normal”.

## VI. RESULTADOS EXPERIMENTALES

### A. Entrenamiento de la CNN

El entrenamiento y la validación de la CNN se realizó utilizando un 4185 imágenes y 1047 imágenes, respectivamente. Debido a la poca cantidad de imágenes y el desbalance de datos se realizó un proceso de *Data Augmentation* mediante el uso de *ImageDataGenerator* de Keras, donde se definieron las siguientes transformaciones:

- **Rotación:** Se rotaron aleatoriamente las imágenes en un rango de 0 a 180 grados, simbolizando las variaciones posibles en la orientación de las estructuras pulmonares en las radiografías.
- **Zoom:** Se aplicó un zoom aleatorio hacia adentro o hacia afuera en las imágenes, imitando posibles variaciones en la escala de las radiografías.
- **Desplazamiento:** Se desplazaron las imágenes de forma horizontal y/o vertical una fracción del ancho total de la imagen, contribuyendo a la variación en la posición de las estructuras pulmonares dentro de la radiografía.
- **Volteo Horizontal:** Se voltearon horizontalmente las imágenes, lo que genera versiones reflejadas de las radiografías.

Todas las transformaciones anteriormente aplicadas buscaban mejorar la generalización del modelo, introduciendo variabilidad a las posiciones de las estructuras en las imágenes. De esta forma el modelo aprende las diferencias intrínsecas de las imágenes pulmonares más que la distribución en las posiciones de estas estructuras.

Para realizar el entrenamiento se eligió el optimizador RMSprop, con función de pérdida de entropía cruzada binaria y basándose como métrica en la *precisión* del conjunto de validación para elegir los mejores pesos del modelo durante el entrenamiento. Durante el entrenamiento se realizaron *checkpoints* de los mejores pesos basándose en la métrica antes mencionada, *early stopping* y *learning rate adaptativo* para intentar obtener los mejores pesos para los conjuntos de entrenamiento y validación, evitando lo mas posible el *overfitting*.

La primera fase de entrenamiento se realiza en 25 épocas, considerando un *batchsize* de 130 imágenes. Los resultados del entrenamiento se reportan en la imagen 7. El *accuracy* alcanzado sobre el conjunto de validación es de 96%.

### B. Métricas de Evaluación del Modelo

Existen variadas métricas que permiten mostrar la capacidad de generalización de un modelo, siendo evaluadas dependiendo de las predicciones del modelo. Estas predicciones se pueden clasificar de la siguiente forma en nuestro modelo:

- True Positive o **TP**: Predicho con neumonía y tiene neumonía en realidad.

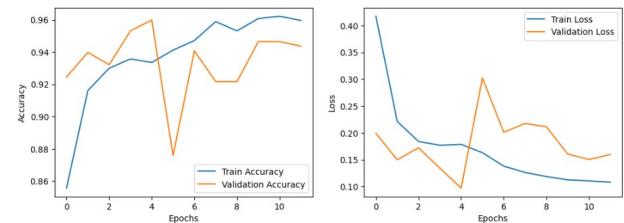


Fig. 7. El gráfico de la izquierda traza el *accuracy* del modelo para el conjunto de entrenamiento y validación, y el gráfico de la derecha representa la función de pérdida sobre el conjunto de entrenamiento y validación. Se efectuaron 25 épocas, quedando la precisión final sobre el conjunto de validación fue de un 96%

- True Negative o **TN**: Predicho normal y normal en realidad.
- False Positive o **FP**: Predicho neumonía y normal en la realidad.
- False Negative o **FN**: Predicho normal y neumonía en la realidad.

a partir de estas clasificaciones se definen muchas métricas. A continuación se realiza el cálculo de algunas de ellas sobre el modelo propuesto:

1) *Matriz de Confusión*: Una matriz de confusión es una representación matricial de los resultados de las predicciones que se utiliza a menudo para describir el rendimiento del modelo de clasificación sobre un conjunto de datos de prueba cuyos valores reales se conocen. En la Figura 8 se aprecia la matriz de confusión resultante del modelo sobre los datos de *Testing*, quedando la clasificación:

- **TP** = 384.
- **TN** = 189
- **FP** = 45
- **FN** = 6

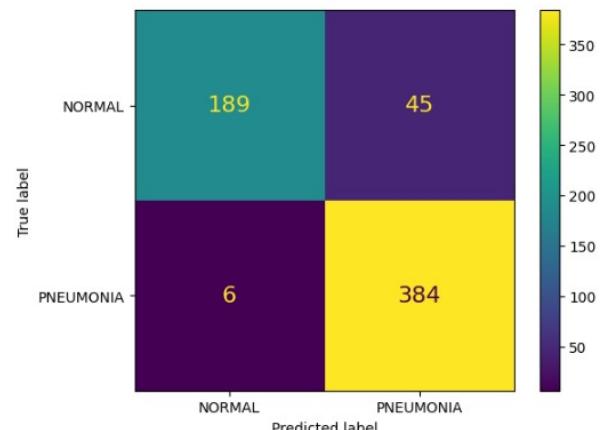


Fig. 8. Matriz de confusión sobre el conjunto de test. Se observa que el modelo tiene un pequeño bias hacia la clase *neumonía*

2) *Exactitud*: La exactitud o accuracy mide el el porcentaje total de elementos clasificados correctamente.

$$\begin{aligned} Accuracy &= \frac{TP + TN}{Total} \\ &= \frac{384 + 189}{624} = 0.91 \end{aligned} \quad (1)$$

3) *Sensibilidad*: La sensibilidad o Recall mide el número de elementos identificados correctamente como positivos del total de positivos verdaderos.

$$\begin{aligned} Recall &= \frac{TP}{TP + FN} \\ &= \frac{384}{384 + 6} = 0.98 \end{aligned} \quad (2)$$

4) *Precisión*: La precisión mide el número de elementos identificados correctamente como positivo de un total de elementos identificados como positivos.

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ &= \frac{384}{384 + 45} = 0.89 \end{aligned} \quad (3)$$

5) *F1-Score*: El F1-Score es una métrica que combina tanto la Precisión como la Sensibilidad en una sola métrica que permite calificar la calidad del modelo predictor.

$$\begin{aligned} F1 - Score &= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \\ &= \frac{20.89 \cdot 0.98}{0.89 + 0.98} = 0.93 \end{aligned} \quad (4)$$

A partir de estas métricas se puede observar una alta sensibilidad del modelo a la neumonía, pero una precisión moderada debido a la cantidad de Falsos positivos. El F1-Score demuestra que el modelo es bastante bueno en su trabajo, pero aún podría mejorar un poco más si aumentamos su precisión.

#### C. Aplicación de los métodos XAI

Para comprender qué áreas de la radiografía de tórax son resaltadas por el modelo CNN para la detección de neumonía y para los casos de pacientes sanos, se seleccionan dos imágenes torácicas del conjunto de test, una correspondientes a un paciente que ha sido diagnosticado sin neumonía (etiqueta “normal”) y que el modelo predice como una imagen de la clase “normal”, y una radiografía torácica asociada a un paciente diagnosticado con neumonía (etiqueta “neumonía”) y que el modelo predice como una imagen perteneciente a la clase “neumonía”. Las imágenes se presentan en la columna (a) de la Figura 2. El modelo entrenado predice que la primera imagen corresponde a un paciente etiquetado como “normal” con score de 0.819. En el segundo caso, la CNN predice con score de 0.928 que la imagen corresponde a un paciente etiquetado con “neumonía”.

En la Figura 2 se presentan los resultados de aplicar el método LIME para explicar la predicción del modelo sobre la clasificación de una imagen torácica asociada a un paciente sano y a un paciente con neumonía. Notamos que en el caso

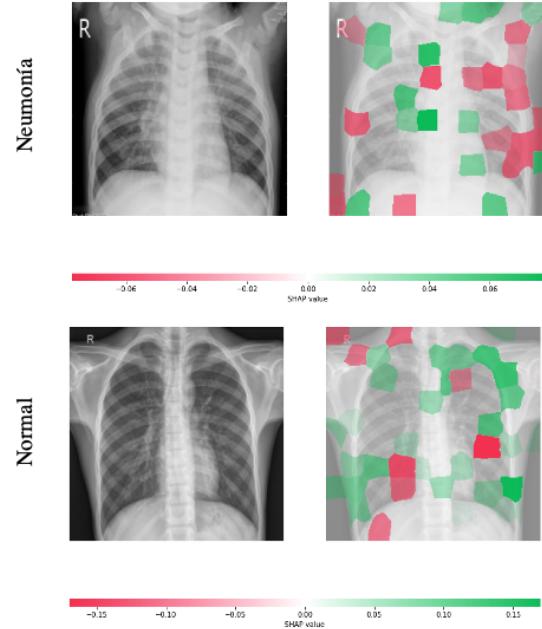


Fig. 9. Resultados aplicando el método Shap. Las imágenes ubicadas a la derecha de cada radiografía representan la atención del modelo para clasificar en cada una de las dos clases.

“neumonía” el modelo le asigna importancia a la zona ubicada entre ambos pulmones.

Por otro lado, en la figura 4 se presentan los resultados de aplicar el método Silency Maps y SmoothGrad. Para el caso neumonía, la atención del modelo está en la zona inferior del pulmón derecho. Destacamos que en el caso “normal” la atención se ubica en el mismo sector que para el caso “neumonía” pero con menor intensidad.

Los resultados obtenidos aplicando el método SHAP se presentan en la Figura 9. Se destaca que la atención del modelo para clasificar la imagen etiquetada como “neumonía” en la clase neumonía, está en la parte superior del pulmón derecho y en el sector central-derecho del pulmón izquierdo, observando tonos verdes oscuros, que significan magnitudes altas en comparación a los otros tres casos.

Finalmente, los mapas de prominencia obtenidos con método RISE son presentado en la Figura 11. Notamos que para el caso del paciente con neumonía, la atención del modelo esta en el sector superior del pulmón derecho, describiendo la imagen en este sector un color rojo intenso. Por otro lado, para el caso del paciente sano, notamos una distribución relativamente simétrica de los colores y zonas pintadas con menor intensidad.

#### D. Score de las métricas de Inserción y Borrado

La Tabla I muestra una evaluación comparativa entre los distintos enfoques de explicabilidad seleccionados, en términos de las métricas de eliminación e inserción. Para cada método se reporta el valor promedio y la desviación estándar alcanzada sobre un conjunto de 15 imágenes seleccionadas al azar del conjunto de test. La métrica reportada corresponde al área bajo la

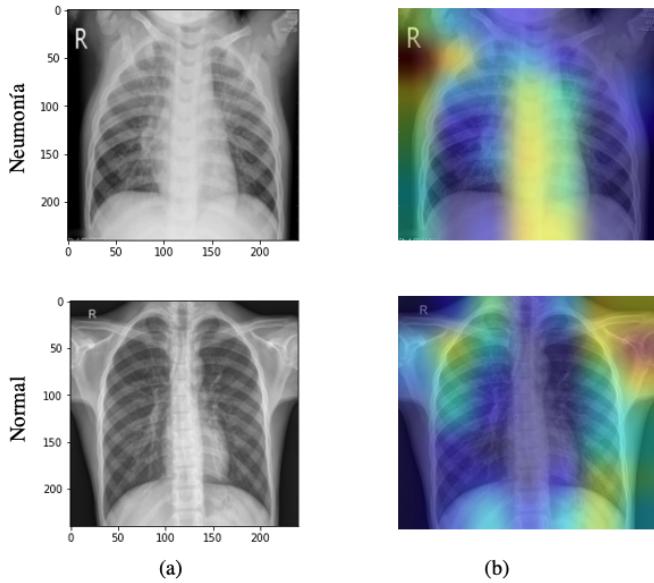


Fig. 10. Resultados aplicando el método Grad CAM. La columna (a) presenta radiografía torácicas asociadas a un paciente con neumonía (arriba) y un paciente sano (abajo). La columna (b) muestra los mapas de prominencia asociados a la predicción del modelo para la clase correspondiente a la etiqueta de cada imagen.

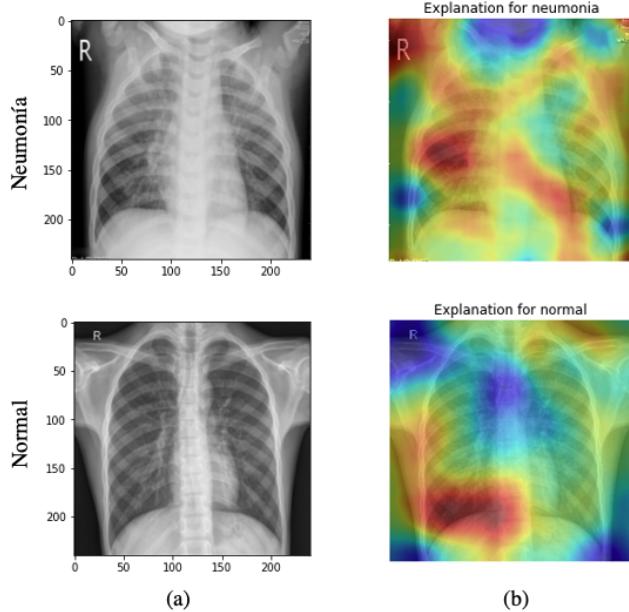


Fig. 11. Resultados aplicando el método RISE. La columna (a) presenta radiografía torácicas asociadas a un paciente con neumonía (arriba) y un paciente sano (abajo). La columna (b) muestra los mapas de prominencia asociados a la predicción del modelo para la clase correspondiente a la etiqueta de cada imagen.

curva de la gráfica que describe la relación entre la probabilidad de clasificación asociada a la clase neumonía y la proporción de pixeles agregados o eliminados, según sea el caso (ver Figuras 5-6).

Según lo señalado en la subsección V-C, un valor alto en la métrica de inserción es bueno y un valor pequeño en la métrica de borrado también es positivo. Para el modelos base propuesto y de acuerdo con ambas métricas, LIME proporciona un mejor rendimiento.

TABLE I  
MÉTRICAS CAUSALES SOBRE IMÁGENES DEL CONJUNTO TEST

Método	Inserción	Borrado
RISE	$0.3399 \pm 0.1256$	$0.3060 \pm 0.1243$
Smooth Grad	$0.7909 \pm 0.1061$	$0.6665 \pm 0.0647$
Silency Map	$0.9259 \pm 0.05165$	$0.8430 \pm 0.0621$
SHAP	$0.3096 \pm 0.1421$	$0.3643 \pm 0.1538$
LIME	$0.6270 \pm 0.2199$	$0.5629 \pm 0.1676$

#### E. Señalamiento Humano

La Figura 12 corresponde a los distintos mapas de prominencia seleccionados y aplicados a una radiografía torácica de un paciente con neumonía bacteriana. La imagen ha sido segmentada por un ingeniero biomédico, marcando las zonas donde existe presencia de inflamación (zonas dentro del pulmón de mayor densidad). La imagen ha sido clasificada por el modelo propuesto como neumonía con probabilidad 0.99. Se desprende de la Figura 12 que existe parcial coincidencia entre el señalamiento humano y los métodos de RISE, SHAP y Smooth Grad. En este caso Grad CAM le asigna importancia a la zona del corazón. Finalmente, Silency Map no aporta un mapa de prominencia que se relacione con el señalamiento humano.

La Figura 13 corresponde a los distintos mapas de prominencia seleccionados y aplicados a una radiografía torácica de un paciente con neumonía viral. La imagen ha sido segmentada por un ingeniero biomédico, marcando las zonas donde existe presencia de inflamación (zonas dentro del pulmón de mayor densidad). La imagen ha sido clasificada por el modelo propuesto como neumonía con probabilidad 0.99. Se desprende de la Figura 13 que existe parcial coincidencia entre el señalamiento humano y los métodos de LIME, RISE, SHAP y Smooth Grad. Por otro lado, también en este caso Grad CAM le asigna importancia a la zona del corazón. Para concluir, Silency Map no aporta un mapa de prominencia que se relacione con el señalamiento humano.

#### VII. CONCLUSIONES Y PROYECCIONES

Se desprende de la matriz de confusión presentada en la Figura 8, y las métricas de evaluación que el modelo entrenado clasifica de manera correcta la clase “neumonía”, pero tiene cierto problema al reconocer las imágenes como “normal”, 8. Lo anterior puede ser debido al gran desbalance entre las clases presente en el conjunto de entrenamiento y/o un modelo que no generaliza lo suficientemente bien, o que el método aplicado para corregir el desbalance no ha sido efectivo. Pese a que el modelo tiene un alto accuracy y recall, tiene una

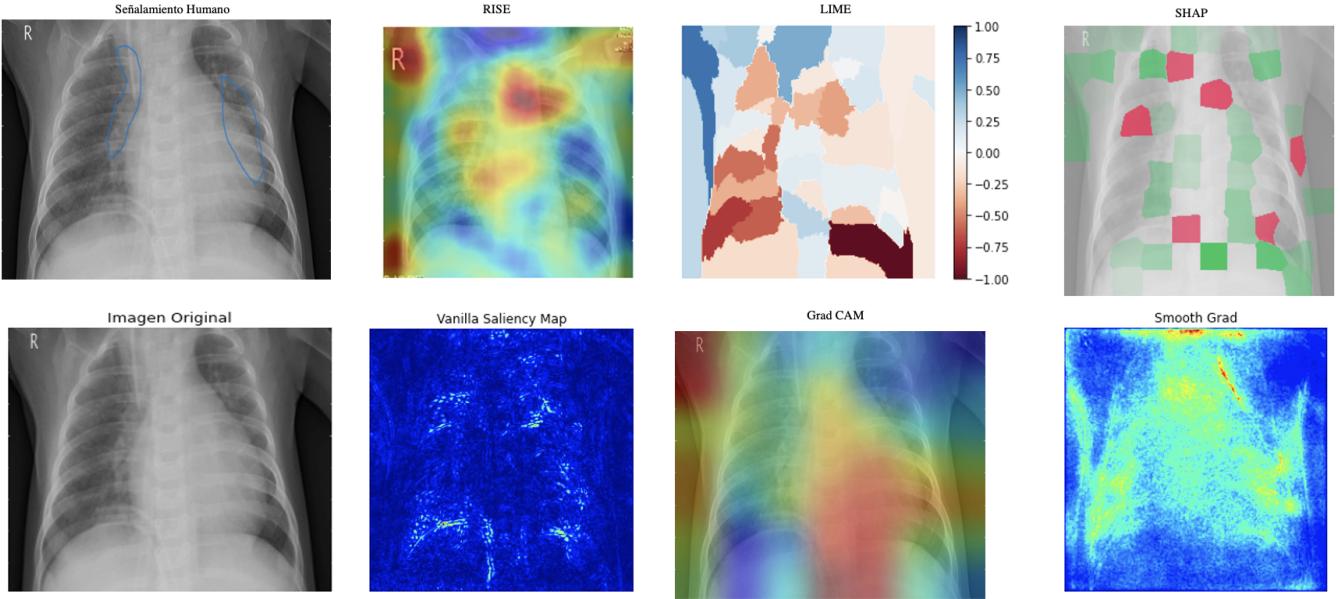


Fig. 12. Señalamiento Humano correspondiente a paciente con neumonía bacteriana.

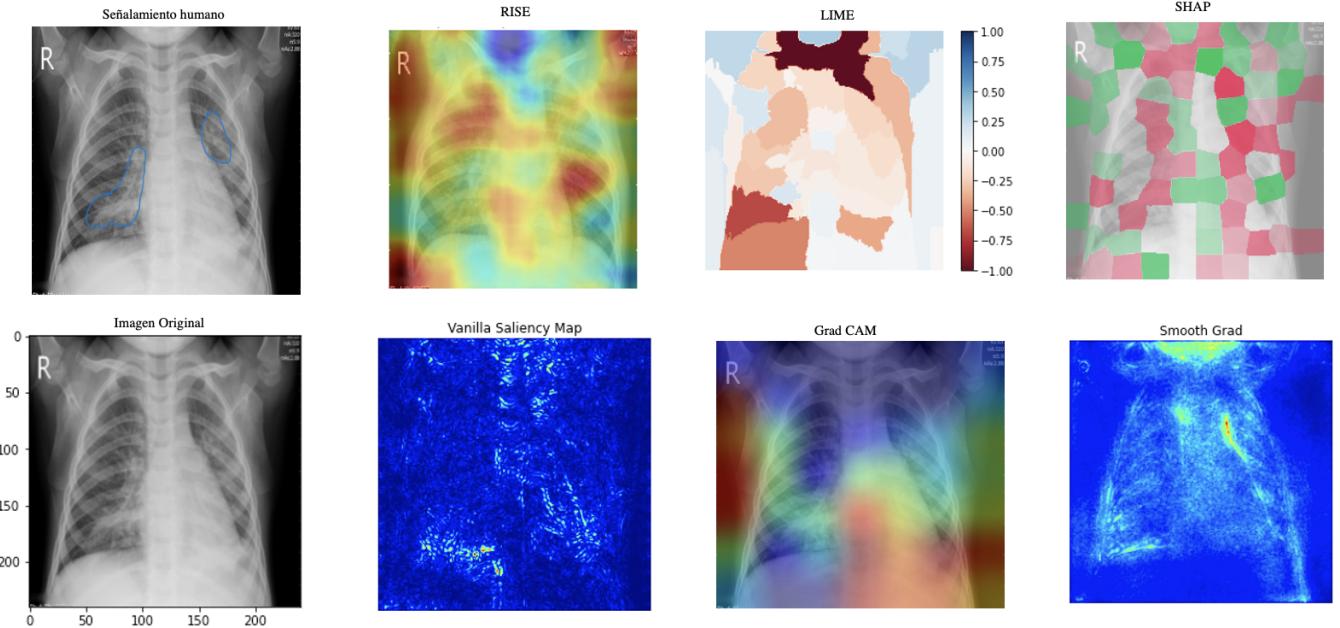


Fig. 13. Señalamiento Humano correspondiente a paciente con neumonía viral.

precisión relativamente lo que es demostrado por el desempeño del modelo con las imágenes normales. Es importante destacar que un mal modelo conlleva a una predicción errónea, por lo que es de vital importancia generar un modelo lo suficientemente preciso y sensible para obtener explicaciones útiles.

Se esperaba que para el caso normal la atención del modelo no tuviese preferencia por una zona en particular, mientras que para los casos de neumonía se esperaba que el modelo destaque las zonas más afectadas por la enfermedad. Sin embargo no

se observó este comportamiento, sino que el modelo tuvo explicaciones erráticas, muchas veces fijándose en partes de la imagen que no pertenecían a los pulmones, como los codos o el cuello, en especial en las imágenes normales, lo que indica que hace falta un método para guiar al modelo a fijarse estrictamente en la zona torácica de la radiografía. Esto permitiría que el modelo se fije en otras características más relevantes para el trabajo en cuestión.

A partir de las explicaciones obtenidas con los métodos,

destacamos que para la imagen etiquetada como neumonía existe cierta similaridad en los resultados obtenidos con LIME, SHAP y RISE. Notamos que los dos casos asociados a Silency map, el modelo no diferencia de manera significativa las zonas de atención asociadas a la predicción del modelo para los casos “neumonía” y “normal”, sino que la atención calculada del modelo es mucho más amplia, destacándose pequeñas zonas con muchísima granularidad, lo que termina afectando en la utilidad de la explicación; además de fijarse muchísimo en el cuello del paciente cuando se usa SmoothGrad.

Por otro lado, se incorporó una métrica de señalamiento humano para las imágenes clasificadas como neumonía. El experto interpretó áreas de mayor densidad, patrones, etc, que permiten diagnosticar esta tipo de enfermedad. De las figuras 12 y 13 se deduce que, para las imágenes seleccionadas, existe un match parcial entre las zonas que marcó el experto y los métodos de explicabilidad seleccionados, destacando en esta correspondencia SHAP, LIME y Smooth grad. Para futuros trabajos, lo anterior requiere cuantificar esta correspondencia midiendo la cantidad de pixeles destacados por el método caen dentro de la región segmentada, como lo presentado en [15].

Finalmente, el presente trabajo requiere implementar métricas automáticas, por ejemplo de inserción y borrado [11], que permitan discriminar que método de explicabilidad resulta mas conveniente para apoyar la predicción que hace el modelo en los casos “neumonía”. Se abre la puerta también a la creación de un modelo muchísimo más robusto y preciso que clasifique con mayor exactitud las radiografías torácicas, para que de esa manera se pueda analizar con mayor profundidad y correctitud el método más óptimo para las explicaciones de clasificación de neumonía en radiografías médicas.

#### REFERENCES

- [1] D. S. Kermany, M. Goldbaum, W. Cai, C. C.S. Valentim, H. Liang, S. L. Baxter, et al., “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.
- [2] R. A. Adegbola, “Childhood pneumonia as a global health priority and the strategic interest of the Bill & Melinda Gates Foundation,” *Clinical Infectious Diseases*, vol. 54, Suppl 2, pp. S89–S92, 2012.
- [3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: explaining the predictions of any classifier. arXiv (Cornell University).
- [4] Kamath, U., & Liu, J. (2021). Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning. Springer International Publishing.
- [5] Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.
- [6] Brunese, L., Mercaldo, F., Reginelli, A., & Santone, A. (2020). Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. *Computer Methods and Programs in Biomedicine*, 196, 105608. <https://doi.org/10.1016/j.cmpb.2020.105608>
- [7] W. Ausawalaithong, A. Thirach, S. Marukatat and T. Wilairasitporn, “Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach,” 2018 11th Biomedical Engineering International Conference (BMEICON), Chiang Mai, Thailand, 2018, pp. 1-5, doi: 10.1109/BMEICON.2018.8609997.
- [8] Rajpal, Sheetal et al. ‘COV-ELM Classifier: An Extreme Learning Machine Based Identification of COVID-19 Using Chest X-ray Images’. 1 Jan. 2022 : 193 – 203.
- [9] Van Der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022b). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79, 102470. <https://doi.org/10.1016/j.media.2022.102470>
- [10] Huang GB, Zhu QY, Siew CK. Extreme learning machine:a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541). Vol. 2. IEEE; 2004. pp. 985–990.
- [11] Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:1806.07421 [cs.CV]. <https://arxiv.org/abs/1806.07421>
- [12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- [13] ImageNet. (n.d.). <https://image-net.org/>
- [14] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
- [15] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Xiaohui Shen Jonathan Brandt, Stan Sclaroff (2017). Top-down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*.