FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TRABAJO DE INVESTIGACIÓN

Título de la Investigación

Comparación de métodos de explicación del comportamiento de modelos de aprendizaje profundo en el procesamiento de imágenes digitales

PARA OPTAR EL GRADO ACADÉMICO DE BACHILLER EN INGENIERÍA DE SISTEMAS

Autor(es)

Jimenez Lucumi Vicenta del Rosario
ORCID: https://orcid.org/0000-0008-0005-0954

Mercado Sarmiento Francois Bernard
ORCID: https://orcid.org/0000-0007-0004-0841

Asesor

Dr. Tuesta Monteza Victor Alexci
ORCID: https://orcid.org/0000-0002-5913-990X

Línea de Investigación

Ciencias de la información como herramientas multidisciplinares y estratégicas en el contexto industrial y de organizaciones

Sublínea de Investigación

Nuevas tendencias digitales orientadas al análisis y uso estratégico de la información

Pimentel – Perú

2024

Jimenez Lucumi

Comparación de métodos de explicación del comportamiento de modelos de aprendizaje profundo en el pr

Universidad Señor de Sipan

Detalles del documento

Identificador de la entrega
trn:oid:::26396:409680470

Fecha de entrega
26 nov 2024, 8:35 a.m. GMT-5

Fecha de descarga
26 nov 2024, 8:36 a.m. GMT-5

Nombre de archivo
tesis turnitin.docx

Tamaño de archivo
3.0 MB

30 Páginas

9,068 Palabras

53,707 Caracteres

9% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto mencionado
- Coincidencias menores (menos de 8 palabras)

Fuentes principales

4%	Fuentes de Internet
0%	Publicaciones
6%	Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Universidad Señor de Sipán

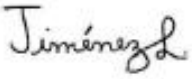
DECLARACIÓN JURADA DE ORIGINALIDAD

Quienes suscriben la DECLARACIÓN JURADA, somos **Jimenez Lucumi Vicenta del Rosario, Mercado Sarmiento Francois Bernard** del Programa de Estudios de **Ingeniería de Sistemas** de la Universidad Señor de Sipán, declaramos bajo juramento que somos autores del trabajo titulado:

COMPARACIÓN DE MÉTODOS DE EXPLICACIÓN DEL COMPORTAMIENTO DE MODELOS DE APRENDIZAJE PROFUNDO EN EL PROCESAMIENTO DE IMÁGENES DIGITALES

El texto de mi trabajo de investigación responde y respeta lo indicado en el Código de Ética de la Universidad Señor de Sipán, conforme a los principios y lineamientos detallados en dicho documento, en relación con las citas y referencias bibliográficas, respetando el derecho de propiedad intelectual, por lo cual informo que la investigación cumple con ser inédito, original y autentico.

En virtud de lo antes mencionado, firman:

Jimenez Lucumi Vicenta del Rosario	DNI: 72072120	
Mercado Sarmiento Francois Bernard	DNI: 75904195	

Pimentel, 17 de julio del 2024.

Dedicatoria

Dedico este logro a Dios, quien ha sido mi guía y mi fortaleza en cada paso de este camino. A mis padres, José Jimenez y Yvonne Lucumi, mi más profundo agradecimiento por su amor, esfuerzo y sacrificio incondicional; su apoyo ha sido la piedra angular de mis logros. A mi hermano Eugenio, gracias por ser mi compañero y mi fuente de motivación. A mis abuelos, por su constante fe en mí y por enseñarme el valor del esfuerzo. Y a mis amigas y familia, por su compañía y apoyo incondicional en cada momento, llenando mis días de risas y aliento. Este logro es el resultado del esfuerzo y apoyo de todos ustedes. ¡Infinitas gracias!

Vicenta del Rosario Jimenez Lucumi

Dedico esta tesis a mis padres, quienes han sido mi pilar fundamental, apoyándome en cada desafío que he enfrentado. A mi familia, por ser la luz que me guía y la motivación constante en mi vida. También quiero expresar mi agradecimiento a mis amigos y seres queridos, por su confianza y aliento incondicional. Este logro es un reflejo del esfuerzo y amor de todos ustedes. ¡Gracias de todo corazón!

Francois Bernard Mercado Sarmiento

Agradecimiento

Queremos expresar nuestro más profundo agradecimiento a la Universidad Señor de Sipán por su apoyo constante y los recursos que facilitaron nuestro desarrollo académico. Asimismo, deseamos reconocer a todos los profesores, cuyas enseñanzas y guía han sido esenciales para nuestro progreso. Un agradecimiento especial a nuestro asesor, Víctor Alexci Tuesta Montenegro, por su valiosa orientación y respaldo durante este proceso. Este logro es, en gran parte, un reflejo del compromiso y la dedicación de todos quienes nos brindaron su apoyo.

Índice

Índice de tablas e ilustraciones

Resumen

Este estudio tuvo como objetivo comparar diferentes métodos de explicación del comportamiento de modelos de aprendizaje profundo en el procesamiento de imágenes digitales. Se implementaron tres métodos ampliamente reconocidos: Grad-CAM, LIME y Occlusion Sensitivity, aplicándolos a tres datasets distintos: Reusimat_USS_Dataset, Brain Tumor MRI Dataset y Pistachio Dataset. Los modelos de aprendizaje profundo utilizados fueron ResNet50, EfficientNetV2B0 y MobileNetV2. La efectividad de los métodos de explicación se evaluó mediante las métricas de fidelidad, monotonía y robustez. Los resultados mostraron que EfficientNetV2B0 alcanzó la mayor precisión (99%) en los datasets de tumores cerebrales y pistachos. Grad-CAM demostró alta monotonía en todos los datasets, mientras que LIME obtuvo la mayor fidelidad en el dataset Reusimat_USS. Occlusion Sensitivity mostró un rendimiento excepcional en los datasets de tumores cerebrales y pistachos, con alta fidelidad y robustez. Se concluyó que la efectividad de los métodos de explicación varía significativamente según el contexto y las características de los datos, subrayando la importancia de un enfoque adaptativo en la selección de métodos de explicación. Este estudio contribuye al avance de la explicabilidad en inteligencia artificial y sienta las bases para futuras investigaciones que busquen equilibrar el rendimiento de los modelos con su interpretabilidad.

Palabras Clave: Aprendizaje profundo, Explicabilidad de IA, Procesamiento de imágenes, Redes neuronales convolucionales, Interpretabilidad de modelos

Abstract

This study aimed to compare different methods for explaining the behavior of deep learning models in digital image processing. Three widely recognized methods were implemented: Grad-CAM, LIME, and Occlusion Sensitivity, applied to three distinct datasets: Reusimat_USS_Dataset, Brain Tumor MRI Dataset, and Pistachio Dataset. The deep learning models used were ResNet50, EfficientNetV2B0, and MobileNetV2. The effectiveness of the explanation methods was evaluated using the metrics of fidelity, monotonicity, and robustness. The results showed that EfficientNetV2B0 achieved the highest accuracy (99%) on the brain tumor and pistachio datasets. Grad-CAM demonstrated high monotonicity across all datasets, while LIME obtained the highest fidelity on the Reusimat_USS dataset. Occlusion Sensitivity showed exceptional performance on the brain tumor and pistachio datasets, with high fidelity and robustness. It was concluded that the effectiveness of the explanation methods varies significantly depending on the context and characteristics of the data, highlighting the importance of an adaptive approach in selecting explanation methods. This study contributes to the advancement of explainability in artificial intelligence and lays the groundwork for future research seeking to balance model performance with interpretability.

Keywords: Deep Learning, AI Explainability, Image Processing, Convolutional Neural Networks (CNNs), Model Interpretability

I. INTRODUCCIÓN

1.1. Realidad problemática

En la era actual de la inteligencia artificial, los modelos de aprendizaje profundo han emergido como una herramienta poderosa que ha transformado diversas aplicaciones, desde el reconocimiento de objetos en imágenes hasta la toma de decisiones autónomas. Impulsados por redes neuronales artificiales y técnicas como las redes neuronales convolucionales (CNN) y los modelos generativos, estos avances han demostrado una precisión que a menudo supera las capacidades humanas [1]. Sin embargo, este progreso está acompañado por desafíos significativos relacionados con la explicabilidad de las decisiones tomadas por estos modelos.

A pesar de estos desafíos, la comunidad de ingeniería de aprendizaje profundo está trabajando activamente en encontrar soluciones efectivas. Se han desarrollado enfoques de interpretación y métodos de explicación de modelos, buscando generar explicaciones comprensibles para los seres humanos [2]. Aunque estos métodos representan un paso hacia la superación de la "caja negra," aún están en desarrollo y presentan limitaciones propias, lo que destaca la necesidad continua de investigar soluciones efectivas que equilibren la precisión del modelo con la capacidad de proporcionar explicaciones claras [3].

El aprendizaje profundo se basa en redes neuronales, computadoras que aprenden y se adaptan a partir de datos. [4] Dentro de este enfoque, destacan diversas arquitecturas de redes neuronales, cada una con sus aplicaciones específicas.

El perceptrón, unidad básica, es útil para clasificaciones simples y sienta las bases para modelos más complejos, como las Redes Neuronales Multicapa (MLP), utilizadas ampliamente en tareas de clasificación y regresión [4].

Las Redes Neuronales Convolucionales (CNN) son esenciales para procesar imágenes y detectar patrones visuales en datos bidimensionales [5]. Las Redes Neuronales Recurrentes (RNN) se especializan en secuencias de datos temporales, como el procesamiento de lenguaje natural [5].

Finalmente, las Redes Adversarias Generativas (GAN) son innovadoras en la creación de datos realistas y se aplican en campos como la generación de imágenes y la síntesis de datos para mejorar conjuntos de entrenamiento [6].

El procesamiento de imágenes es un campo crucial en la visión por computadora, donde diversas técnicas son empleadas para preparar, describir y analizar imágenes digitales. En este contexto, el preprocesamiento de imágenes juega un papel fundamental, y abarca procesos como el redimensionado y la normalización, que ajustan las

dimensiones y escalas de las imágenes, respectivamente, para optimizar su uso en algoritmos y aplicaciones específicas [7]. Además, el aumento de datos, una técnica clave, implica la generación de nuevas instancias de datos mediante transformaciones variadas y aleatorias, contribuyendo así a la diversificación y generalización de los modelos de aprendizaje automático.

La extracción y selección de características son etapas esenciales en el procesamiento de imágenes, donde se identifican y utilizan atributos representativos para describir eficientemente las propiedades visuales de una imagen. Desde histogramas de color hasta descriptores locales, una variedad de técnicas son empleadas para este propósito, con el objetivo de reducir la dimensionalidad de los datos y mejorar la eficiencia computacional.

En cuanto a la segmentación de imágenes,[8] esta técnica divide una imagen en partes o regiones significativas basadas en ciertas características visuales, facilitando así el análisis y la comprensión de su contenido. Métodos como la umbralización y la segmentación por regiones son comúnmente utilizados para este fin, con aplicaciones en campos tan diversos como la medicina y la navegación autónoma.

El avance del aprendizaje profundo ha revolucionado el procesamiento de imágenes, permitiendo a las computadoras interpretar, analizar y comprender el contenido visual con una precisión asombrosa. Desde la clasificación de imágenes hasta la generación de imágenes realistas, las aplicaciones del aprendizaje profundo en imágenes son diversas y ampliamente utilizadas en la actualidad.

A pesar del éxito alcanzado en tareas de procesamiento de imágenes digitales, la falta de capacidad para explicar cómo llegan a sus conclusiones o decisiones los modelos de aprendizaje profundo plantea una serie de desafíos en el campo de la ingeniería de sistemas [9]. En primer lugar, esta falta de transparencia dificulta la identificación de posibles fallos o sesgos en su funcionamiento [10]. Además, esta carencia de explicación puede tener implicaciones éticas y legales significativas, especialmente en aplicaciones críticas como diagnósticos médicos asistidos por inteligencia artificial [11]. Aunque estos modelos han demostrado una precisión impresionante, su reputación de "caja negra" surge del hecho de que no pueden proporcionar explicaciones razonables para las decisiones que toman.

Un ejemplo concreto de esta problemática se encuentra en la detección de anomalías en imágenes médicas. Aunque los sistemas basados en aprendizaje profundo han avanzado significativamente en esta área, su incapacidad para proporcionar una explicación clara sobre por qué fallan cuando cometen errores sigue siendo un desafío importante [12].

El objetivo de esta investigación es comparar diferentes métodos de explicación del comportamiento de modelos de aprendizaje profundo en el procesamiento de imágenes digitales. Se busca identificar las fortalezas y debilidades de cada método, proporcionando explicaciones claras y comprensibles sobre los resultados generados por el aprendizaje profundo aplicado a imágenes digitales. Este documento aborda la evaluación de tres métodos ampliamente reconocidos: Grad-CAM, LIME y Occlusion Sensitivity. Además, se detalla el proceso de aplicación de estos métodos a tres conjuntos de datos específicos y se presentan los resultados obtenidos a través de su implementación.

1.2. Formulación del problema

¿Cuál es el método de explicación más efectivo para modelos de aprendizaje profundo en el procesamiento de imágenes digitales?

1.3. Hipótesis

La comparación entre los métodos de explicación permitirá identificar el método más efectivo para explicar modelos de aprendizaje profundo en el procesamiento de imágenes digitales.

1.4. Objetivos

Objetivo general

Comparar diferentes métodos de explicación del comportamiento de modelos de aprendizaje profundo en el procesamiento de imágenes digitales

Objetivos específicos

- Identificar los métodos de explicación del comportamiento de modelos de aprendizaje profundo propuestos en la literatura reciente.
- Seleccionar los métodos de explicación más relevantes y adecuados para la comparación.
- Implementar los métodos de explicación seleccionados en un entorno de procesamiento de imágenes digital.
- Comparar la efectividad de los métodos de explicación en la interpretación del comportamiento de modelos de aprendizaje profundo.

1.5. Teorías relacionadas al tema

[13] propusieron el método PRISM (Principal Image Sections Mapping) para visualizar representaciones internas de CNNs. Utiliza PCA para mapear características

detectadas por la CNN en una máscara RGB. Experimentos revelaron que PRISM es fiel, con un cambio promedio absoluto de 0.54 en la clasificación al intercambiar partes de representaciones entre imágenes. Aplicable a diversas CNNs, PRISM es ligeramente más rápido que Grad-CAM. Destaca características comunes y distintivas en imágenes de perros border collie y husky siberiano. Limitaciones incluyen la falta de indicación de importancia para clasificación final y menor confiabilidad con baja varianza en los 3 primeros componentes principales.

[14] presentaron el método "perceptual quality-preserving" (PQP) para generar ataques adversarios efectivos en clasificadores de imágenes basados en redes neuronales profundas en un escenario de caja negra. PQP inyecta ruido adversarial en regiones "seguras" de la imagen, identificadas por un bajo gradiente local del índice de similitud estructural (SSIM), manteniendo la calidad perceptual. En experimentos, PQP superó consistentemente a métodos de referencia. En CIFAR-10, logró una tasa de éxito del 98%, con solo 1593 consultas en promedio, manteniendo un SSIM de 0.989. En reconocimiento facial MCS2018, obtuvo hasta un 95.8% de éxito con 13400 consultas y SSIM de 0.978. Limitaciones incluyen el uso de imágenes pequeñas y la falta de exploración de ataques dirigidos a clases específicas.

[15] llevaron a cabo una revisión exhaustiva de métodos de mejora de imágenes con poca luz basados en deep learning, comparando enfoques CNN y GAN. Las CNN utilizan capas convolucionales para aprender características y mejorar la relación entre imágenes de entrada y salida. Los métodos GAN emplean un generador y un discriminador compitiendo por producir imágenes realistas. SCI-easy destacó con los mejores puntajes en métricas como PSNR (29.46 dB), SSIM (0.8776), y NIQE (2.78) en el conjunto de datos NPE, superando a Zero-DCE y CERL. Aunque Zero-DCE mostró los mejores efectos visuales subjetivos, algunos métodos se centran en mejorar el rendimiento sin considerar la generalización en aplicaciones del mundo real. Además, la dependencia de grandes conjuntos de datos de entrenamiento plantea desafíos.

[16] presentaron Slideflow, un conjunto de herramientas de aprendizaje profundo para histopatología digital. Slideflow procesa eficientemente imágenes histopatológicas completas extrayendo azulejos a diferentes niveles de magnificación, almacenados en formatos compatibles con TensorFlow y PyTorch. Incluye métodos como aprendizaje supervisado débil, aprendizaje de instancias múltiples, GANs, cuantificación de incertidumbre y explicabilidad del modelo. En una evaluación sobre carcinoma de células escamosas de cabeza y cuello, un modelo de clasificación basado en azulejos supervisado débilmente logró un AUC de 0.87 y una precisión promedio de 0.80 en un

conjunto de prueba externo. Con la cuantificación de incertidumbre, el 89.5% de las diapositivas tuvieron predicciones de alta confianza, con un AUC de 0.88 y una precisión del 90.2%. Limitaciones incluyen dependencia de backends específicos y la ausencia de ciertas funciones, como el análisis de gráficos de núcleos con redes neuronales de grafos.

[17] presentaron un marco de trabajo interpretable para el análisis y planificación del tratamiento de la enfermedad de Alzheimer (EA) basado en datos multimodales. Emplearon un enfoque de múltiples etapas que integraba datos tabulares, imágenes, texto y expresión génica, utilizando modelos de IA y técnicas de Inteligencia Artificial Explicable (XAI). En la identificación de áreas infectadas en imágenes de resonancia magnética (RM), VGG16 superó a la CNN convencional con un área bajo la curva ROC de 0.98 frente a 0.96. Factores importantes para pacientes con y sin demencia fueron identificados, y regiones cerebrales relevantes fueron identificadas en imágenes de RM. Los genes asociados con la EA también fueron identificados. Aunque no se mencionaron limitaciones específicas, posibles desafíos incluyen la disponibilidad y calidad de los datos multimodales y la necesidad de validación en conjuntos de datos más grandes y diversos.

[18] introdujeron un método innovador para predecir la susceptibilidad de los incendios forestales, empleando un modelo de aprendizaje profundo basado en un transformer codificador, complementado con la técnica SHAP para mejorar la interpretabilidad. Adaptaron la arquitectura transformer para procesar datos numéricos no secuenciales y optimizaron los hiperparámetros con PSO. El modelo alcanzó alta precisión, con un RMSE de 0.12 en validación cruzada de 10 pliegues, superando a otros modelos como SVM y ANN. El análisis de SHAP reveló la velocidad del viento como el factor más influyente (SHAP promedio de 0.14). Sin embargo, el rendimiento del modelo disminuyó en un conjunto de datos de evaluación independiente, destacando la necesidad de interpretabilidad proporcionada por SHAP. La limitación incluye la generalización del modelo a otros ecosistemas y regiones.

[19] presentaron MHCXAI, un marco de trabajo que utiliza técnicas de Inteligencia Artificial Explicable (XAI) como LIME y SHAP para generar explicaciones interpretables de los predictores de presentación de péptidos en el Complejo Principal de Histocompatibilidad (CMH) de clase I basados en aprendizaje profundo. Se generaron perturbaciones en la secuencia de péptidos de entrada, se predijo la salida para estas instancias perturbadas y se ajustó un modelo interpretable simple. Tanto LIME como SHAP produjeron explicaciones válidas y consistentes, con LIME exhibiendo mayor

estabilidad y consistencia, mientras que SHAP fue más preciso en la evaluación de la importancia de los aminoácidos y posiciones en el péptido-CMH. Los coeficientes de correlación de Pearson entre las explicaciones de LIME y SHAP y el valor ΔΔG de BAlaS fueron principalmente positivos. Sin embargo, la evaluación se basó en aproximaciones como ΔΔG de BAlaS, lo que limita la validación.

[20] introdujeron un marco de trabajo para detectar carcinoma de células escamosas (SCC) en múltiples órganos mediante un modelo generalizado. Preprocesaron imágenes histopatológicas y utilizaron una técnica de selección de características combinando tres métodos. Para la clasificación, emplearon CatBoost con ajuste de hiperparámetros de Optuna. El modelo alcanzó una precisión del 96,66% y un coeficiente de correlación de Matthews (MCC) de 0,9334 en su conjunto de datos privado, y una precisión del 91,45% y un MCC de 0,8305 incluso con variaciones de iluminación. Superó a modelos de aprendizaje profundo preentrenados, como ResNet50 y EfficientNet-B0. Incorporaron técnicas explicables de aprendizaje automático, como ELI5, LIME y SHAP, para brindar interpretabilidad. Sin embargo, las limitaciones incluyen el tamaño del conjunto de datos y la falta de abordar otros desafíos comunes en imágenes histopatológicas.

[21] presentaron un marco de aprendizaje profundo basado en un mecanismo de atención para clasificar la retinopatía diabética. Descompusieron la imagen de retina para separar la atención en estructuras oscuras y brillantes, generando mapas de atención interpretables. Emplearon la arquitectura Xception y la función de pérdida focal, y aplicaron el mecanismo de atención a las estructuras de la retina. Lograron una precisión del 83,7% y un Kappa Ponderado Cuadrático de 0,78 en la clasificación de 5 grados de severidad en el conjunto de prueba de Kaggle DR. Aunque comparables a otros enfoques, destacaron por los mapas de atención interpretativos. Sin embargo, presentaron limitaciones como el algoritmo de descomposición previa es lento, y tiende a fallar en la clasificación de ciertos grados de severidad y a subdiagnosticar otros. Además, los mapas de atención tienen baja resolución, dificultando la detección precisa de lesiones retinianas.

[22] propusieron el método ABELE para mejorar la confianza y comprensión de los sistemas de diagnóstico de lesiones cutáneas basados en IA. ABELE utiliza un autoencodador adversario progresivo para generar imágenes sintéticas coherentes con los datos originales, actuando como ejemplares y contra-ejemplares. Estas explicaciones visuales junto con un mapa de saliencia resaltan las regiones clave para la clasificación. En la encuesta de validación, los participantes con formación médica

lograron una precisión del 91,26% utilizando las explicaciones de ABELE. Se observó un aumento significativo en la confianza, especialmente cuando el modelo estaba inicialmente equivocado. Sin embargo, también se encontró una disminución en la confianza entre los expertos después de recibir consejos incorrectos. Las limitaciones incluyen la dependencia de un conjunto de datos predefinido que puede introducir sesgos y el enfoque centrado en profesionales médicos, con menos atención a la comprensión de los pacientes.

[23] presentaron NeuroNet19, un nuevo modelo de red neuronal profunda para la clasificación de tumores cerebrales en imágenes de resonancia magnética. Integraron la arquitectura VGG19 con un innovador Módulo de Agrupación de Pirámide Invertida (iPPM) para mejorar la extracción de características a múltiples escalas. Tras el preprocesamiento de las imágenes, NeuroNet19 utilizó VGG19 como columna vertebral para extraer características, las cuales se refinaron mediante el módulo iPPM y capas adicionales de convolución y agrupación global. Los resultados fueron destacables, con NeuroNet19 logrando una precisión del 99.3%, precisión del 99.2%, recuperación del 99.2%, puntaje F1 del 99.2% y coeficiente Kappa de Cohen del 99%. Estas cifras superaron a otros modelos como ResNet50, VGG16, MobileNet y DenseNet121. Sin embargo, el modelo se entrenó solo en cuatro tipos de tumores específicos, lo que podría limitar su aplicabilidad en nuevos casos, y se reconoció la necesidad de ampliar el conjunto de datos para mejorar su versatilidad clínica.

[24] propusieron un modelo de aprendizaje profundo con inteligencia artificial explicable (XAI) para detectar cálculos renales en imágenes de rayos X KUB. Modificaron la arquitectura preentrenada VGG16 y aplicaron la propagación de relevancia por capas (LRP) como técnica de XAI. El modelo entrenado logró una impresionante precisión de prueba del 97,41%, con una tasa de error del 2,59%, una precisión del 97,39%, una sensibilidad del 98,45%, una especificidad del 95,70% y un puntaje F1 de 0,98 en un conjunto de prueba de 4279 imágenes de rayos X KUB. Las limitaciones incluyen la disponibilidad de datos de alta calidad y diversos de imágenes médicas de rayos X KUB de cálculos renales, así como la necesidad de mejorar la interpretación del modelo a pesar del uso de XAI.

[25] propusieron un sistema de inteligencia artificial explicable (XAI) para diagnosticar melanomas, basado en una ontología dermatológica y datos anotados por expertos. Utilizaron ResNet50 para desarrollar un clasificador de imágenes dermoscópicas y un esquema de explicación multimodal. El sistema XAI alcanzó una precisión equilibrada del 81% en el conjunto de prueba. Las explicaciones del sistema estuvieron altamente

alineadas con las de los dermatólogos, con un solapamiento del 0.46 en las explicaciones ontológicas y del 0.48 en las regiones de interés. Aunque el sistema XAI no mejoró la precisión diagnóstica de los clínicos, aumentó significativamente su confianza en sus diagnósticos propios y en el sistema de soporte. Sin embargo, el estudio se realizó en condiciones artificiales y no consideró el cambio de dominio, lo que limita su aplicabilidad en entornos clínicos reales y otros contextos sin ontologías específicas.

[26] propusieron aplicar técnicas de Inteligencia Artificial Explicable (XAI) para interpretar modelos de aprendizaje profundo en la clasificación de enfermedades hepáticas. Utilizaron el enfoque SHAP para generar explicaciones sobre las predicciones de sus modelos de redes neuronales profundas. El método implicó preprocesamiento de datos, entrenamiento de modelos de clasificación con TensorFlow, y aplicación del paquete SHAP para interpretar las predicciones. El mejor modelo alcanzó una precisión del 82%, precisión del 72%, recall del 81%, y medida F de 0.76. Los gráficos SHAP mostraron la influencia de cada característica en las predicciones, facilitando la comprensión de tratamientos personalizados. Las limitaciones incluyeron la necesidad de reproducir el estudio en otros contextos médicos y explorar enfoques alternativos de XAI, así como desarrollar marcos específicos para la implementación en el sector de la salud.

[27] propusieron un modelo de red neuronal convolucional (CNN) liviano y optimizado para la clasificación en tiempo real de imágenes de resonancia magnética cardíaca (CMR) de pacientes con enfermedad arterial coronaria (CAD) y sujetos sanos. Adaptaron la arquitectura LeNET-5 y realizaron un ajuste exhaustivo de hiperparámetros. Implementaron la CNN en el conjunto de datos CMR de CAD y aplicaron técnicas de Inteligencia Artificial Explicable (XAI) como GradCAM y Layer-wise Relevance Propagation (LRP) para generar mapas de calor que resaltaran las regiones relevantes. El modelo logró una precisión de clasificación del 99.35% y una precisión balanceada del 99.13%. En la validación cruzada estratificada de 10 pliegues, mantuvo una precisión similar del 99.22% y una precisión balanceada del 99.10%. Sin embargo, el análisis por fotograma y la dependencia del contraste y la calidad de las imágenes fueron limitaciones identificadas.

[28] propusieron desarrollar un modelo de inteligencia artificial explicable para analizar imágenes de heridas vasculares en pacientes asiáticos. Utilizaron un conjunto de 2957 imágenes de un hospital terciario de Singapur para entrenar el modelo. Emplearon modelos convolucionales preentrenados para la clasificación de tipos de heridas,

medición de profundidad, ancho y largo, así como para la segmentación de heridas. El modelo logró una precisión promedio del 95.9% para la clasificación de heridas, con un AUROC de 0.99. Para la medición de profundidad, alcanzó una precisión del 85.0% y un AUROC de 0.97. En cuanto al ancho y largo, obtuvo una precisión del 87.1% y un AUROC de 0.92, mientras que para la segmentación de heridas logró una precisión del 87.8% y un AUROC de 0.95. Sin embargo, la insuficiencia de datos en algunas categorías de heridas y los desafíos en la interpretación de la explicabilidad de la IA fueron limitaciones identificadas.

[29] propusieron un enfoque de aprendizaje profundo llamado "DNet-SVM" para la detección temprana de enfermedades en plantas de caña de azúcar. El método utilizó una combinación de la arquitectura DenseNet201 como generador de características y un clasificador de máquina de vectores de soporte (SVM). Además, implementaron el modelo de interpretabilidad LIME para explicar las predicciones del modelo. DNet-SVM superó a modelos de transferencia de aprendizaje convencionales, logrando una sensibilidad de 0.94, una especificidad de 0.86, una tasa de falsos negativos de 0.05 y una tasa de falsos positivos de 0.13, con una precisión de clasificación del 97%. Aunque el enfoque mostró resultados prometedores, las limitaciones potenciales incluyen la necesidad de una gran cantidad de datos de entrenamiento y desafíos en la interpretabilidad del modelo, especialmente con el enfoque de aproximación de LIME.

[30] propusieron un sistema de apoyo a la toma de decisiones clínicas (CDSS) basado en inteligencia artificial explicable para la Enfermedad Pulmonar Obstructiva Crónica (EPOC), utilizando el modelo ontológico TMR y la argumentación. Implementaron el sistema como microservicios integrados en un sistema de registro electrónico de salud (EHR) en tiempo real. La evaluación del sistema mostró un alto acuerdo con los neumólogos (97% en la asignación del grupo GOLD) y opiniones favorables sobre su implementación. En términos numéricos, el CDSS sugirió la terapia más favorecida en el 31.3% de los casos, con tasas de selección del 33.3%, 24.2%, y 9.1% para las siguientes opciones. Sin embargo, la pandemia de COVID-19 retrasó la evaluación en Serbia y algunas limitaciones prácticas, como la necesidad de ingreso manual de datos y restricciones del proveedor de EHR, afectaron la implementación y evaluación del sistema.

[31] propusieron un enfoque de aprendizaje profundo para el diagnóstico del cáncer de ovario, combinando un modelo de red neuronal convolucional ensamblado con cuatro ramas y un modelo de transformador para segmentación de imágenes. Las imágenes segmentadas alimentaron la arquitectura CNN, donde cada rama realizó extracción de

características individualmente. El modelo alcanzó una precisión del 98,96% en la clasificación de tumores ováricos, superando a clasificadores individuales. Además, el modelo de transformador superó a U-Net en segmentación, con puntajes de Dice de 0,98 y 0,99 para imágenes benignas y malignas, respectivamente. Se implementó un modelo de conjunto de aprendizaje automático con datos de biomarcadores clínicos, logrando una precisión del 92,85%. Sin embargo, la generalización del modelo puede ser limitada por el tamaño del conjunto de datos y se requieren validaciones clínicas rigurosas para su aplicación en entornos reales.

[32] propusieron evaluar el rendimiento de métodos de Inteligencia Artificial Explicable (XAI) como Grad-CAM, Grad-CAM++, y Eigen-CAM en la detección de cáncer de mama en mamografías. Utilizaron el puntaje del "Juego de Puntería" para comparar los mapas de saliencia generados por estos métodos con las anotaciones de radiólogos expertos. Entrenaron un modelo de aprendizaje profundo ResNet50 en la detección de cáncer de mama y luego aplicaron los métodos de XAI para resaltar regiones relevantes. Los resultados mostraron que los puntajes del Juego de Puntería fueron 0.41, 0.30, y 0.35 para Grad-CAM, Grad-CAM++, y Eigen-CAM, respectivamente, en el conjunto de prueba. Incluso en casos verdaderos positivos, los puntajes aumentaron marginalmente. Las limitaciones incluyen un tamaño de muestra pequeño, la exclusión de otras técnicas de XAI, y el remuestreo que podría haber afectado la calidad de los mapas de saliencia y su evaluación.

[33] propusieron Contrastive Layer-wise Relevance Propagation (CLRP) como un método para generar explicaciones discriminativas a nivel de clase para las decisiones de clasificación de redes neuronales convolucionales profundas (DCNNs). CLRP, basado en Layer-wise Relevance Propagation (LRP), modela un concepto visual opuesto para contrastar con la clase objetivo, redistribuyendo la puntuación de la clase objetivo a otras clases o negando los pesos de la última capa conectada. Los resultados en conjuntos de datos de ImageNet mostraron que CLRP supera a LRP en generación de explicaciones. En el juego de señalamiento, CLRP1 y CLRP2 superaron a LRP con una precisión de 38,44% y 39,13% respectivamente para VGG16. En un estudio de ablación, CLRP1 y CLRP2 también superaron a LRP con puntuaciones promedio de activación de neurona caídas de 0,2093 y 0,2030 respectivamente para AlexNet. Sin embargo, CLRP todavía enfrenta desafíos en la discriminación fina dentro de clases similares.

[34] presentaron el "Swap Test" (ST) para explicar decisiones de modelos de aprendizaje profundo en el diagnóstico de la enfermedad de Alzheimer con imágenes

de resonancia magnética cerebral. El ST intercambia parches de una imagen con imágenes de referencia opuestas, obteniendo un mapa de calor de regiones relevantes para el diagnóstico. Los resultados mostraron que el ST exhibe continuidad y selectividad, con valores de continuidad significativamente más bajos (por ejemplo, 15.492 vs 30.596 de línea base). Además, el ST mostró una fuerte correlación negativa para la selectividad en todos los modelos (por ejemplo, -0.643 para VGG16 2D+C). Una limitación es la incapacidad para establecer relaciones causales, requiriendo más estudios sobre la relación entre marcadores neurobiológicos y explicaciones del ST.

[35] propusieron validar métodos de Inteligencia Artificial Explicable (XAI) para modelos de predicción del éxito estudiantil en el aprendizaje en línea y aulas invertidas. Utilizaron redes neuronales BiLSTM en nueve cursos diferentes y aplicaron los métodos LIME y SHAP para explicabilidad. Los resultados cuantitativos mostraron que LIME y SHAP no coincidían en la importancia de características para un curso individual. Comparando pares de cursos, se encontraron patrones como la relevancia del comportamiento inicial en los MOOC y de la interacción en las aulas invertidas. Cualitativamente, el 85.71% de los educadores generaron ideas accionables a partir de las explicaciones, pero no mostraron preferencia por un método en particular. Las limitaciones incluyen el tamaño limitado del estudio y la necesidad de información demográfica de los estudiantes.

[36] introdujeron un método de Explicación de Inteligencia Artificial (XAI) para la detección de enfermedades en hojas de papa mediante redes neuronales convolucionales profundas. Su enfoque implica perturbar la imagen de entrada con máscaras generadas iterativamente en función de los resultados intermedios del modelo. Compara propuestas de cajas delimitadoras generadas con la predicción objetivo y analiza las regiones de interpretación. Los resultados muestran que su método supera a D-RISE en las métricas de eliminación e inserción. Por ejemplo, para el modelo 2, su método logró puntajes de eliminación de 0.114 y de inserción de 0.697, superando a D-RISE con 0.275 y 0.532, respectivamente. Sin embargo, las limitaciones incluyen la necesidad de probar el método en otros modelos de detección y enfermedades de plantas, así como la falta de mención de limitaciones específicas del método propuesto.

[37] propusieron aplicar algoritmos de Inteligencia Artificial Explicable (XAI) para mejorar la transparencia en la clasificación de cambios pulmonares en pacientes con fibrosis quística utilizando resonancias magnéticas (MRI). Revisaron algoritmos XAI y seleccionaron cuatro: Integrated Gradients, Class Activation Maps, Grad CAM y Layer-wise Relevance Propagation. Estos se aplicaron a una tubería de aprendizaje profundo para clasificar defectos de perfusión pulmonar y tapones de moco. Los resultados

mostraron que los algoritmos XAI generaron mapas de atribución en segundos, destacando áreas relevantes. El 20,14% de las MRI clasificadas en la categoría 1 y el 45,54% en la categoría 4 para defectos de perfusión pulmonar con CAM. Sin embargo, GCam tuvo resultados inferiores, lo que requiere más investigación. Las limitaciones incluyen la degradación de la calidad de imagen y la dependencia de la tubería de clasificación. En general, los algoritmos XAI ofrecieron información útil para los radiólogos, resaltando áreas relevantes para la clasificación.

[38] propusieron una revisión exhaustiva de seis técnicas de Inteligencia Artificial Explicable (XAI) modelo-agnóstico en un caso de estudio sobre cáncer de mama. Utilizaron un modelo de Random Forest para clasificación y aplicaron técnicas como PDP, PFI, ICE, LIME y SHAP para explicar el modelo de caja negra. Las seis técnicas identificaron las características más influyentes, como Survival Months y Regional Node Positive, mientras que características menos significativas incluyeron Estrogen Status y A Stage. Para la explicación global, PDP, PFI y SHAP coincidieron en las características críticas como Survival Months y Age. Sin embargo, hubo divergencias entre el modelo original y el sustituto, con una divergencia del 0.55, sugiriendo la falta de precisión del modelo sustituto. Además, se observaron inestabilidades menores en PFI y LIME debido a perturbaciones aleatorias de datos. Estos hallazgos destacan la necesidad de mejorar la estabilidad y convergencia en XAI.

[39] propusieron examinar el impacto de las explicaciones visuales en la confianza de los usuarios en las clasificaciones de inteligencia artificial. Utilizaron métodos como LIME, Grad-CAM y SHAP en el conjunto de datos ImageNet. Realizaron un estudio con usuarios, entrevistas en profundidad y análisis de oclusión. Los resultados mostraron que las explicaciones visuales fueron más útiles cuando los usuarios tenían conocimiento previo sobre la clasificación. La probabilidad de clase afectó significativamente la confianza de los usuarios, y diferentes métodos de explicación mostraron incoherencias. Los modelos preentrenados también demostraron fragilidad. ResNet152 tuvo una precisión del 82.3% y del 96.0% para las precisiones@1 y@5, respectivamente. Sin embargo, las explicaciones visuales aún eran susceptibles a ajustes subjetivos, siendo una limitación clave del estudio.

[40] propusieron el método McXai para mejorar la confiabilidad de los modelos de caja negra al explicar las decisiones detrás de ellos. Utilizando la búsqueda de árbol de Monte Carlo, formalizaron el problema como dos juegos: uno para encontrar características esenciales y otro para identificar características sensibles a errores. En comparación con métodos post-hoc como LIME y SHAP, McXai necesitó menos pasos

para cambiar la predicción del modelo, con promedios de 4.82, 7.23 y 6.23 pasos respectivamente en el conjunto de datos MNIST. Además, eliminar características negativas encontradas por McXai mejoró la precisión del modelo, por ejemplo, de 93.97% a 94.58% para GoogleNet. Las limitaciones incluyen el tiempo de ejecución influenciado por la complejidad del modelo y la falta de abordaje directo a la interpretabilidad del modelo subyacente.

[41] propusieron una evaluación centrada en el usuario de métodos de Explicación de Inteligencia Artificial (XAI) para el reconocimiento de imágenes naturales. Diseñaron un experimento en línea donde los participantes identificaban objetos en imágenes reveladas gradualmente según los métodos de XAI como Occlusion, Layer-wise Relevance Propagation (LRP) y Prototypical Part Network (ProtoPNet), además de mapas de importancia generados por humanos (ClickMe). Los resultados mostraron que LRP alcanzó una tasa de reconocimiento del 84%, seguido por ProtoPNet con un 88%. Estos métodos fueron comparables a los mapas generados por humanos (81%). Sin embargo, la correlación entre los mapas de XAI y ClickMe fue débil o nula, indicando diferencias en el razonamiento humano. Las limitaciones incluyeron la exclusión de la relevancia negativa y la restricción del porcentaje de características relevantes, lo que podría ocultar comportamientos no deseados del modelo.

[42] presentaron un marco de evaluación específico de dominio para aumentar la transparencia de los modelos de aprendizaje profundo en el reconocimiento facial. Utilizaron la propagación de relevancia por capas (LRP) para generar mapas de calor y delimitaron las regiones faciales relevantes para cada Unidad de Acción (AU) mediante polígonos definidos por puntos de referencia y conocimiento experto. Los resultados mostraron una correlación entre altos valores de la métrica de evaluación (μ_poly) y un mejor rendimiento de clasificación para algunas AUs en el modelo VGG-16, pero algunas AUs no identificaron las regiones esperadas como importantes. En el modelo ResNet-18, los valores μ_poly fueron más bajos que los valores μ_box, destacando la importancia de definir límites precisos. Una limitación fue la falta de normalización de los valores μ según el tamaño de las regiones, lo que sugiere la necesidad de ponderación respecto a la distribución general de relevancia en las imágenes.

II. MÉTODO DE INVESTIGACIÓN

El proceso de comparar métodos de explicación se ilustra en el diagrama siguiente.

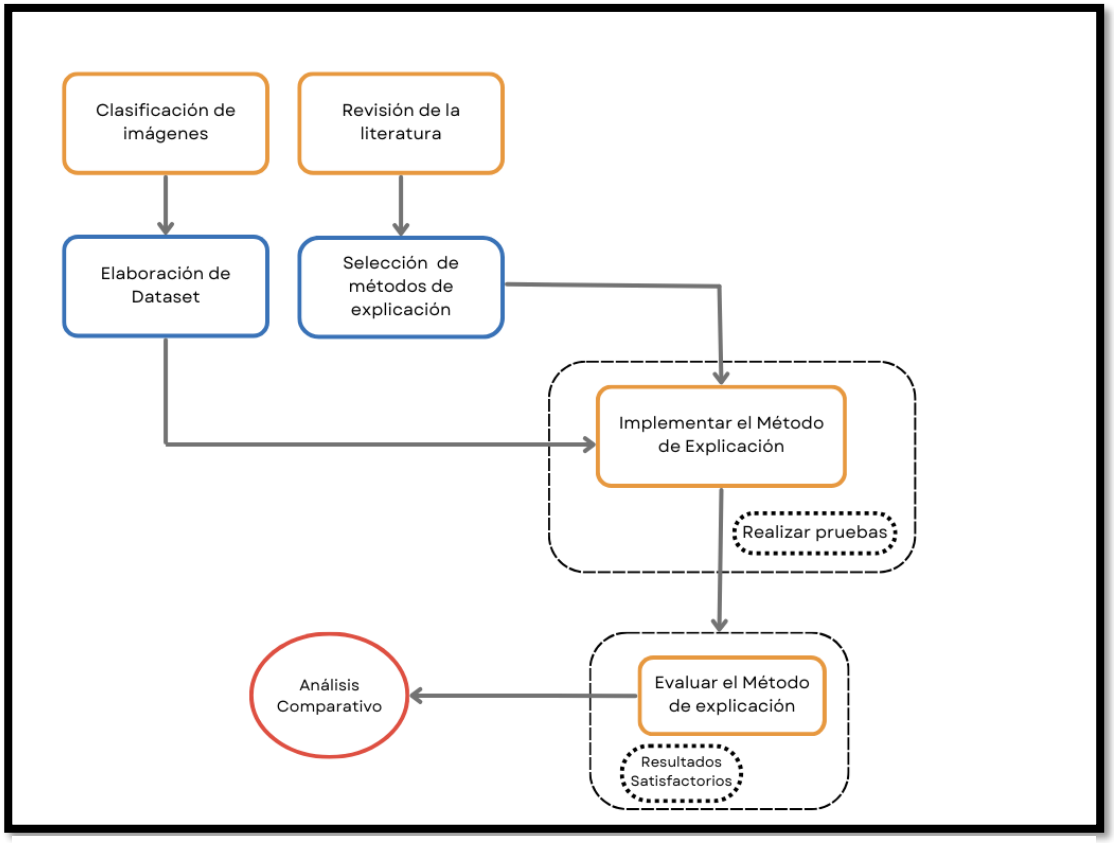


Ilustración 1 Explicación del desarrollo del método

- **Descripción de Dataset**

El dataset utilizado para esta investigación fue proporcionado por los autores José María Torres Chirinos y Firuz Aguilar Casusol. Después de algunas reuniones, se recibió toda la información necesaria para el análisis y uso en los métodos de explicación. El conjunto de datos principal, denominado Reusimat_USS_Dataset, está almacenado en Google Drive y ha sido cuidadosamente organizado para facilitar el entrenamiento y la validación de algoritmos de aprendizaje profundo en el reconocimiento y clasificación de residuos sólidos reutilizables. En total, el conjunto de datos incluye 356 imágenes de artículos reciclados, distribuidas de manera equilibrada entre las diferentes categorías gracias a una técnica de aumento de datos, esencial debido a la cantidad inicial limitada de imágenes disponibles. Esta técnica incluyó normalización de imágenes y transformaciones como corte, zoom y volteo horizontal, mejorando así la generalización del modelo al aumentar la diversidad del conjunto de datos de entrenamiento. El dataset se dividió en dos subconjuntos principales: un 80% para el entrenamiento y un 20% para la validación, asegurando una evaluación

adecuada del modelo en datos no vistos durante el entrenamiento.

Además del Reusimat_USS_Dataset, se utilizaron otros dos conjuntos de datos proporcionados por Kaggle: el Brain Tumor MRI Dataset [43] y el Pistachio Dataset [44]. Estos datasets adicionales se emplearon para reforzar el trabajo con más pruebas y validaciones. El Brain Tumor MRI Dataset se utilizó para experimentar con la clasificación de imágenes médicas, mientras que el Pistachio Dataset se empleó para el reconocimiento y clasificación de diferentes variedades de pistachos.

Para la clasificación de los materiales reciclables en sus respectivas categorías, se utilizaron tres arquitecturas de redes neuronales convolucionales: ResNet50, EfficientNetV2B0 y MobileNetV2. ResNet50 se seleccionó por su capacidad para entrenar redes profundas sin sufrir problemas de desvanecimiento del gradiente, gracias a sus conexiones residuales. EfficientNetV2B0 ofrece un balance óptimo entre precisión y eficiencia computacional mediante un escalado compuesto que ajusta el tamaño de la red. MobileNetV2 fue optimizada mediante capas personalizadas que incluyen agrupación global promedio, capas densas con activación ReLU y dropout para mejorar la generalización y prevenir el sobreajuste, principalmente para el Reusimat_USS_Dataset. En cambio, para los datasets de Brain Tumor MRI y Pistachio, se enfocó más en la utilización de EfficientNetV2B0, debido a su balance entre precisión y eficiencia computacional.

En cuanto a la división de datos, el Reusimat_USS_Dataset y el Pistachio Dataset se dividieron en dos subconjuntos principales: un 80% para el entrenamiento y un 20% para la validación. Para el Brain Tumor MRI Dataset, la división de los datos fue variante entre esos porcentajes, ajustándose a las necesidades específicas del análisis. Este enfoque permite una clasificación efectiva de los materiales y otros objetos en sus categorías correspondientes, proporcionando una base sólida para el entrenamiento de modelos de redes neuronales convolucionales (CNN).

	Contenedor	Material	Nomenclatura
DATAS SET	Black	Papel encerado	waxpaper_001.jpg.
		colillas de cigarro	cigarette_001.jpg.
		residuos sanitarios	sanitarywaste_001.jpg
	Brown	Desechos de alimentos	foodwaste_001.jpg.
		Restos de Poda	Pruning_001.jpg
	Green	Papel	paper_001.jpg
		Cartón	Cardboard_001.jpg.
		Vidrio	Glass_001.jpg …
		plástico	plastic_001.jpg.
		textiles	textiles_001.jpg.
		madera	wood_001.jpg.
		cuero	leather_001.jpg
		empaques compuestos	packaging_001.jpg.
		metales	metals_001.jpg.
	Red	Pilas	Batteries_001.jpg.
		lámparas	lamps_001.jpg …
		luminarias	luminaries_001.jpg.
		medicinas vencidas	medicine_001.jpg.
		envases de plaguicidas, etc.	pesticide_001.jpg

Ilustración 2 Distribución de datos del dataset Reusimat_USS_Dataset

- **Revisión de Literatura**

Para la revisión de literatura, se buscaron artículos detallados en las bases de datos Scopus, ScienceDirect e IEEE Xplore. El objetivo era identificar y desarrollar métodos analíticos para describir métodos de explicación aplicados a imágenes digitales. Los métodos seleccionados para este estudio fueron Grad-Cam, Saliency Maps, DeepLIFT, LIME, SHAP y T-SNE. Para cada uno de estos métodos, se seleccionaron tres casos representativos, excepto para Grad-Cam, del cual se seleccionaron cuatro casos, arrojando un total de 19 casos incluidos en la matriz de análisis.

Se formuló una matriz de análisis para capturar puntos clave de cada caso. Para cada enfoque interpretativo se registran los siguientes elementos: el método específico utilizado, los resultados de las métricas evaluadas y una descripción del contexto y aplicación del método. Este marco permitió una comparación clara y sistemática de los métodos revisados. A continuación, se presentan las matrices de análisis para cada método seleccionado, proporcionando una visión detallada y estructurada de los hallazgos en cada caso.

- **Grad-Cam**

Método	Ref.	Resultados	Descripción
Gradient Mapping Guided Explainable Network (GMGENet)[45]	[1, 2, 3, 4, 5, 6]	Precisión: 0.902 (90.2%) [6] AUC: 0.911 (91.1%) [6] Sensibilidad: 0.909 (90.9%) [6] Especificidad: 0.895 (89.5%) [6]	GMGENet utiliza Grad-CAM para guiar la red neuronal profunda a concentrarse en las regiones relevantes para la identificación de extensión extracapsular (ECE) en imágenes de CT 3D, extrayendo volúmenes de interés (VOIs) sin anotación manual.
Xception con Grad-CAM para BCI[46]	[1, 2, 3, 4, 5, 6]	Precisión: 98.92% [6] Sensibilidad: 99.09% [6] Especificidad: 98.18% [6] F1 Score: 98.91% [6]	Xception emplea transfer learning sobre el dataset de MRI cerebral, utilizando Grad-CAM para generar heatmaps que resaltan las características responsables de las predicciones. El modelo mejora la transparencia en las decisiones de la BCI.
ResNet50 con Grad-CAM[47]	[22, 23, 24, 25]	Precisión: 98.52% [25] Sensibilidad: 97.8% [25] Especificidad: 99.1% [25] AUC: 0.993 [25] F1 Score: 0.981 [25]	Utiliza la arquitectura ResNet50 para detectar tumores cerebrales en imágenes de MRI. Grad-CAM se emplea para generar mapas de activación que destacan las regiones de interés en las imágenes, proporcionando interpretabilidad al modelo. La combinación de ResNet50 y Grad-CAM mejora la detección de tumores cerebrales mediante la integración de técnicas avanzadas de aprendizaje profundo y visualización.
Grad-CAM (Gradient-weighted Class Activation Mapping)[48]	[59, 53, 52, 51, 33]	Precisión en localización: 56.51% (VGG-16) [59] AUC en segmentación débil: 0.496 [59] Correlación con mapas de oclusión: 0.261 [59] Mejora sobre c-MWP: 10.28% [59]	Grad-CAM utiliza gradientes del concepto objetivo que fluyen en la última capa convolucional para producir un mapa de localización que resalta las regiones importantes en la imagen para predecir el concepto. Se combina con Guided Grad-CAM para crear visualizaciones de alta resolución y discriminativas de clase, aplicables a varios modelos CNN sin necesidad de cambios arquitectónicos o reentrenamiento.

Tabla 1 Matriz de artículos del método de explicación Grad -Cam

- **Saliency Maps**

Método	Ref	Resultados	Descripción
Convolutional Neural Networks (InceptionV3, Inception-ResNetV2, VGG-16) con Grad-CAM[49]	[23, 24, 25, 26, 27]	Inception-ResNetV2 Precisión: 95.18% [27] Sensibilidad: 90.35% [27] Especificidad: 100% [27] AUC: 0.918 (95% CI 0.873–0.963) [27] F1 Score: 96.6% [27]	Se utilizan las arquitecturas InceptionV3, Inception-ResNetV2 y VGG-16 preentrenadas con ImageNet y ajustadas con el dataset específico para diferenciar entre tumores renales benignos y malignos. Grad-CAM se emplea para crear mapas de saliencia que destacan las características relevantes en las imágenes CT. El modelo Inception-ResNetV2 mostró el mejor desempeño, enfocándose en la interfaz entre el tumor y el parénquima renal circundante, mejorando así la precisión en la clasificación.
Ensemble de Redes Neuronales Profundas (DNNs) con Saliency Maps[50]	[19, 20, 22, 25, 27]	AUC para no-ERM: 0.99 [27] AUC para pequeño-ERM: 0.92 [27] AUC para grande-ERM: 0.99 [27] Precisión 3-way: 89.33% [27] Sensibilidad: 95.45% [27]]	El estudio utiliza un ensamble de redes neuronales profundas (DNNs) con las arquitecturas ResNet50 e InceptionV3 para detectar y clasificar membranas epirretinales (ERM) en imágenes de OCT. Los mapas de saliencia generados con Guided-Backprop resaltan áreas importantes en las imágenes OCT para una interpretación clara de las decisiones del modelo.
Generación Progresiva de Mapas de Saliencia Model-Agnostic (MAPSM)[51]	[27, 28, 29, 30, 31]	Deletion Score: 0.045 [31] Insertion Score: 0.818 [31] Mean Saliency: 0.085 [31] Saliency Average Contribution (SAC): 2.39 [31]	MAPSM es un método de generación de mapas de saliencia basado en un marco jerárquico para modelos de detección de objetos. A diferencia de otros métodos de caja negra, MAPSM introduce una partición adaptativa de máscaras y una estrategia de generación de máscaras impulsada por la saliencia para reducir el ruido en los mapas de saliencia. Progresivamente descubre y refina las áreas de saliencia de los objetos, resultando en mapas de saliencia más interpretables y de mejor calidad.

Tabla 2 Matriz de artículos del método de explicación Saliency Maps

- **LIME**

Método	Ref	Resultados	Descripción
Shallow-CNN con CAM y LIME[52]	[23, 24, 6, 13, 22]	Precisión: 95% [13] F1 Score: 0.95 [13] AUC: 0.976 [13] Sensibilidad: 0.95 [13] Especificidad: 0.95 [13]	Se desarrolla un Shallow-CNN con cuatro capas de convolución y pooling, optimizado mediante la técnica de optimización bayesiana, para la detección de tuberculosis en radiografías de tórax. Utiliza Class Activation Maps (CAM) y Local Interpretable Model-agnostic Explanations (LIME) para mejorar la interpretabilidad del modelo, mostrando las áreas de interés en las imágenes que contribuyen a la clasificación.
XGBoost con SHAP y LIME[53]	[10, 14, 15, 16, 17]	AUROC: 0.828, 0.913, 0.923 [10] AUPRC: 0.807, 0.796, 0.921 [10] Precisión: 0.785, 0.885, 0.891 [10] F1 Score: 0.63, 0.69, 0.70 [10]	XGBoost se utiliza para predecir la mortalidad a 28 días en pacientes con coagulopatía inducida por sepsis (SIC) en las bases de datos MIMIC-III, MIMIC-IV y eICU-CRD. SHAP y LIME se aplican para proporcionar interpretaciones de las predicciones, destacando la importancia de características como la puntuación SOFA, RDW, y edad en el modelo.
XGBoost con SHAP y LIME[54]	[31, 32, 33, 36, 37]	Precisión: 93.29% [37] Sensibilidad: 91.80% [37] Especificidad: 94.73% [37] AUC: 0.9689 [37] F1 Score: 0.9313 [37]	XGBoost se utiliza para predecir la enfermedad renal crónica (CKD) utilizando características clínicas y de laboratorio. SHAP y LIME se emplean para explicar la influencia de las características en las predicciones del modelo, proporcionando una comprensión detallada de cómo las características individuales afectan los resultados del modelo.

Tabla 3 Matriz de artículos del método de explicación LIME

- **SHAP**

Método	Ref	Resultados	Descripción
Random Forest (RF) con SHAP[55]	[29, 30, 31, 32, 33]	Precisión: 89.66% [33] Sensibilidad: 89.47% [33] Especificidad: 90.00% [33] AUC: 0.9421 [33] F1 Score: 91.89% [33]	Se utiliza el algoritmo Random Forest para construir modelos de predicción utilizando datos de microbiota intestinal y metabolitos. SHAP se aplica para proporcionar explicaciones interpretables a nivel global y personalizado, permitiendo entender la influencia de características individuales en las predicciones del modelo.
XGBoost con SHAP[56]	[33, 34, 35, 36, 37]	AUC: 0.81 [37] Precisión: 0.79 [37] Sensibilidad: 0.62 [37] Especificidad: 0.83 [37] F1 Score: 0.55 [37]	Se utiliza XGBoost para predecir la probabilidad de diagnóstico de síndrome de circulación posterior (PCS) utilizando datos clínicos. SHAP se emplea para interpretar las predicciones del modelo, identificando las características más influyentes en la clasificación de PCS, como el IMC, glucosa en sangre, ataxia, disartria, presión arterial diastólica y temperatura corporal.
Random Forest con SHAP[57]	[23, 24, 25, 26, 27]	AUC: 0.937 (95% CI 0.844–1.000) [26] Sensibilidad: 0.870 [26] Especificidad: 0.900 [26]	Se utiliza el algoritmo Random Forest para construir modelos predictivos utilizando los indicadores genéticos NFKBIA, BCL2A1, y CCL4 para la identificación de la constitución Yin-deficiencia (YinDC). SHAP se aplica para proporcionar explicaciones interpretables, permitiendo visualizar la contribución de cada predictor a las predicciones del modelo y facilitando la comprensión de los resultados del modelo.

Tabla 4 Matriz de artículos del método de explicación SHAP

- **T-SNE**

Método	Ref	Resultados	Descripción
Multimodal Deep Learning con ViViT y Transformer[58]	[16, 17, 18, 19, 22]	Precisión: 88.7% [22] AUC: 0.91 [22] F1 Score: 0.89 [22] Sensibilidad: 87.4% [22] Especificidad: 90.2% [22]	Se utiliza un enfoque de aprendizaje profundo multimodal que combina datos de video de IVIS y parámetros 0D de plasma para predecir las interrupciones en KSTAR. ViViT maneja los datos de video y Transformer maneja los datos 0D. GradCAM y Attention Rollout se utilizan para evaluar la capacidad de los modelos para capturar características relevantes para la predicción de interrupciones.
Red Neuronal con TSNE y Red Neuronal Unicapa[59]	[24, 25, 26, 27, 28]	Precisión: 98.9% [28] Discernibilidad: 98.9% [28] AUC: >0.999 [28] Temporal resolution: 4.5 ms [26] Sensibilidad: 11.4 (mv/Kpa) [26]	Se desarrolló una red neuronal combinada con TSNE para visualizar la diferencia entre imágenes de distintas formas en la matriz sensora de presión piezorresistiva (PRSA) y una red neuronal unicapa para cuantificar la discernibilidad entre las imágenes de diferentes formas. El sistema de sensores muestra una alta resolución espaciotemporal y excelente desempeño en la visualización en tiempo real de múltiples puntos de contacto y seguimiento de la trayectoria del movimiento.
Random Forest (RF) con LASSO y qRT-PCR[60]	[19, 20, 21, 22, 23]	Precisión: 93.50% [23] Sensibilidad: 94.12% [23] Especificidad: 92.85% [23] AUC: 0.967 (95% CI 0.930–0.990) [23]	Utiliza Random Forest combinado con LASSO para identificar genes biomarcadores clave en la osteoporosis postmenopáusica (PMOP). La validación experimental se realiza con qRT-PCR en muestras de sangre de pacientes con PMOP y controles sanos. Este enfoque permite identificar genes clave relacionados con células inmunes que son discriminativos entre altos y bajos niveles de densidad mineral ósea.

Tabla 5 Matriz de artículos del método de explicación T-SNE

- **DeepLIFT**

Método	Ref	Resultados	Descripción
CNN con Saliency Mapping, Guided Grad-CAM y DeepSHAP[61]	[10, 11, 12, 13, 16]	Precisión: 99.33% [16] Sensibilidad: 98.76% [16] Especificidad: 100% [16] AUC: 0.997 [16]	Se utiliza una red neuronal convolucional (CNN) para diagnosticar 10 estados anormales en plantas nucleares. Para aumentar la transparencia del modelo, se aplican tres técnicas explicativas: saliency mapping, Guided Grad-CAM, y DeepSHAP. Estas técnicas identifican los parámetros de entrada más influyentes en la clasificación, optimizando la cantidad de parámetros a monitorear para el diagnóstico eficiente de estados anormales.
Meta-Learning con DeepLIFT [62]	[1, 2, 3, 5, 10]	C-Index (Individual Datasets): 0.74 (Transcriptómica) [3], 0.75 (Clínica) [3], 0.58 (Proteómica) [3] C-Index (Combinaciones): 0.79 (Integrado) [3], 0.84 (Clínica + Transcriptómica) [3], 0.63 (Proteómica + Transcriptómica) [3] Integrated Brier Score: 0.12 [3] Enriquecimiento de vías: DNA repair pathways (P < 0.05) [5]	Utiliza un modelo de meta-aprendizaje combinado con DeepLIFT para el análisis de supervivencia basado en datos multi-ómicos de cáncer. El modelo mejora el análisis de supervivencia al integrar transcriptómica, proteómica y datos clínicos de TCGA, con interpretabilidad proporcionada por DeepLIFT, que asigna puntuaciones de contribución a los genes para identificar vías moleculares relevantes.
Random Forest, LASSO, y DeepLIFT[63]	[54, 55, 56, 57]	Precisión: 0.785 (Random Forest) [56] AUROC: 0.781 (LASSO) [56] F1 Score: 0.758 (DeepLIFT) [56] Validación Cruzada: 0.792 (Promedio) [56]	Se utilizan Random Forest, LASSO y DeepLIFT para predecir la ocurrencia de convulsiones electroencefalográficas (ES) en niños críticamente enfermos. Random Forest construye un modelo a partir de múltiples árboles de decisión, LASSO aplica regularización para seleccionar variables clave, y DeepLIFT asigna puntuaciones de importancia a las características basadas en redes neuronales, permitiendo interpretar la contribución de cada característica en la predicción.

Tabla 6 Matriz de artículos del método de explicación DeepLIFT

III. RESULTADOS

Hemos seleccionado tres métodos de explicación del comportamiento de modelos de aprendizaje profundo en el procesamiento de imágenes digitales: Grad-CAM, LIME y Occlusion Sensitivity. La elección de estos métodos se debe a su amplia adopción en la comunidad de investigación y su capacidad para proporcionar explicaciones visuales e interpretables sobre las predicciones de modelos complejos. Grad-CAM es conocido por su capacidad para generar mapas de calor que resaltan las regiones importantes de la imagen que influyen en la predicción del modelo. LIME ofrece una aproximación local para interpretar modelos de caja negra mediante la generación de explicaciones lineales simples alrededor de cada predicción. Occlusion Sensitivity, por otro lado, utiliza la técnica de oclusión para evaluar la importancia de diferentes partes de la imagen, proporcionando una perspectiva adicional sobre cómo el modelo interpreta los datos.

Para evaluar estos métodos, hemos utilizado los tres conjuntos de datos previamente mencionados: Reusimat_USS_Dataset, Brain Tumor MRI Dataset y Pistachio Dataset. El Reusimat_USS_Dataset está dividido en cuatro categorías de residuos sólidos reutilizables (negro, marrón, verde y rojo). El dataset de Brain Tumor MRI clasifica tres tipos de tumores cerebrales (glioma, meningioma y pituitaria) y la ausencia de tumor. El Pistachio Dataset presenta dos características remarcadas de pistachos, Kirmizi y Siirt.

Los modelos de aprendizaje profundo utilizados en esta evaluación fueron ResNet50, EfficientNetV2B0 y MobileNetV2. A continuación se presenta una tabla con los resultados de las métricas de evaluación (Accuracy, Precision, Recall y F1-Score) para cada combinación de modelo y dataset.

Modelo	Dataset	Accuracy	Precision	Recall	F1-Score
ResNet50	Reusimat_USS_Dataset	0.94	0.94	0.92	0.93
	Brain Tumor MRI	0.97	0.96	0.97	0.96
	Pistachio	0.97	0.96	0.98	0.97

Tabla 7 Métricas de entrenamiento de los dataset con el Modelo ResNet50

Modelo	Dataset	Accuracy	Precision	Recall	F1-Score
EfficientNetV2B0	Reusimat_USS_Dataset	0.98	0.99	0.99	0.98
	Brain Tumor MRI	0.99	0.99	0.99	0.99
	Pistachio	0.99	0.98	1	0.99

Tabla 8 Métricas de entrenamiento de los dataset con el Modelo EfficientNetV2B0

Modelo	Dataset	Accuracy	Precision	Recall	F1-Score
MobileNetV2	Reusimat_USS_Dataset	0.95	0.95	0.94	0.94
	Brain Tumor MRI	0.96	0.95	0.96	0.95
	Pistachio	0.96	0.95	0.97	0.96

Tabla 9 Métricas de entrenamiento de los dataset con el Modelo MobileNetV2

Después de haber realizado una evaluación exhaustiva de los modelos de aprendizaje profundo ya mencionados, se procederá a seleccionar los modelos que hayan obtenido los puntajes más altos. Posteriormente, se aplicarán los métodos de explicación del comportamiento de los modelos: Grad-CAM, LIME y Occlusion Sensitivity.

Para evaluar la efectividad de estos métodos de explicación, utilizamos tres métricas: fidelidad, monotonía y robustez. Estas métricas cuantitativas nos permiten medir la explicabilidad de los modelos y son ampliamente reconocidas en la comunidad de investigación. "Aunque en última instancia es el consumidor quien determina la calidad de una explicación, la comunidad de investigación ha propuesto métricas cuantitativas como indicadores de la explicabilidad" [64]. La fidelidad evalúa la correlación entre la importancia asignada por el algoritmo de interpretabilidad a los atributos y el efecto de cada uno de los atributos en el rendimiento del modelo predictivo. La monotonía mide si la importancia de los atributos asignados por el método de explicación sigue un patrón creciente en relación con el rendimiento del modelo. La robustez, aunque no es mencionada específicamente en AI Explainability 360, es crucial para asegurar que las explicaciones generadas sean consistentes y fiables, independientemente de pequeñas variaciones en los datos de entrada.

Estas evaluaciones nos permitirán entender mejor la calidad y utilidad de las explicaciones generadas por Grad-CAM, LIME y Occlusion Sensitivity, proporcionando una base sólida para la interpretación de los modelos de aprendizaje

profundo en el reconocimiento y clasificación de residuos sólidos reutilizables.

En las tablas siguientes, se presentan los resultados de estas mediciones, permitiéndonos comparar y contrastar la eficacia de cada método de explicación aplicado a los diferentes modelos y datasets utilizados en esta investigación.

Método 1: Grad-Cam

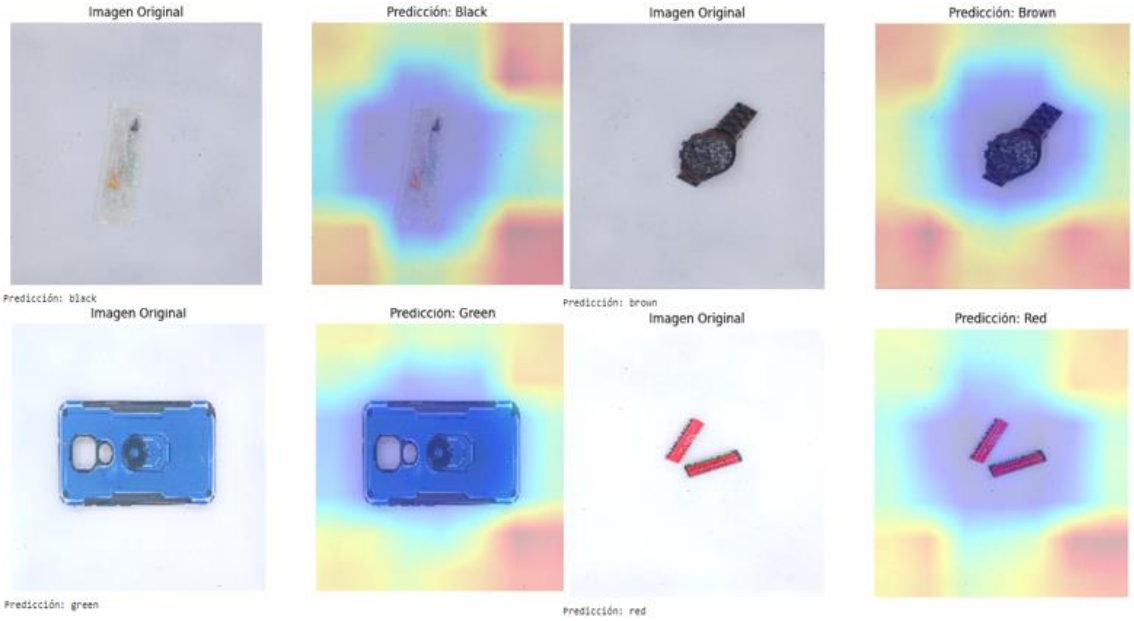


Ilustración 3 Grad-Cam aplicado al Dataset de Reusimat_USS_Dataset

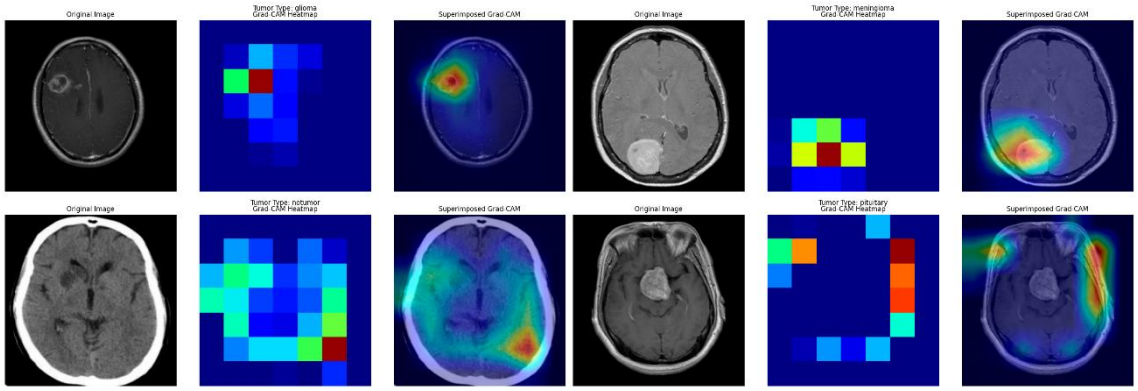


Ilustración 4 Grad-Cam aplicado al Dataset de Brain Tumor MRI

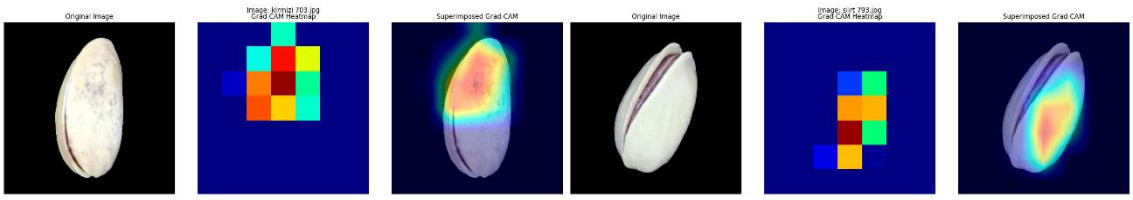


Ilustración 5 Grad-Cam aplicado al Dataset de Pistachio

Dataset	Fidelidad	Monotonía	Robustez
Reusimat_USS_Dataset	0.3296	1.0000	0.7298
Brain Tumor MRI	0.1072	1.0000	0.8275
Pistachio	0.1467	1.0000	0.8022

Tabla 10 Resultados de las métricas aplicadas a los datasets con Grad-Cam

Método 2: LIME

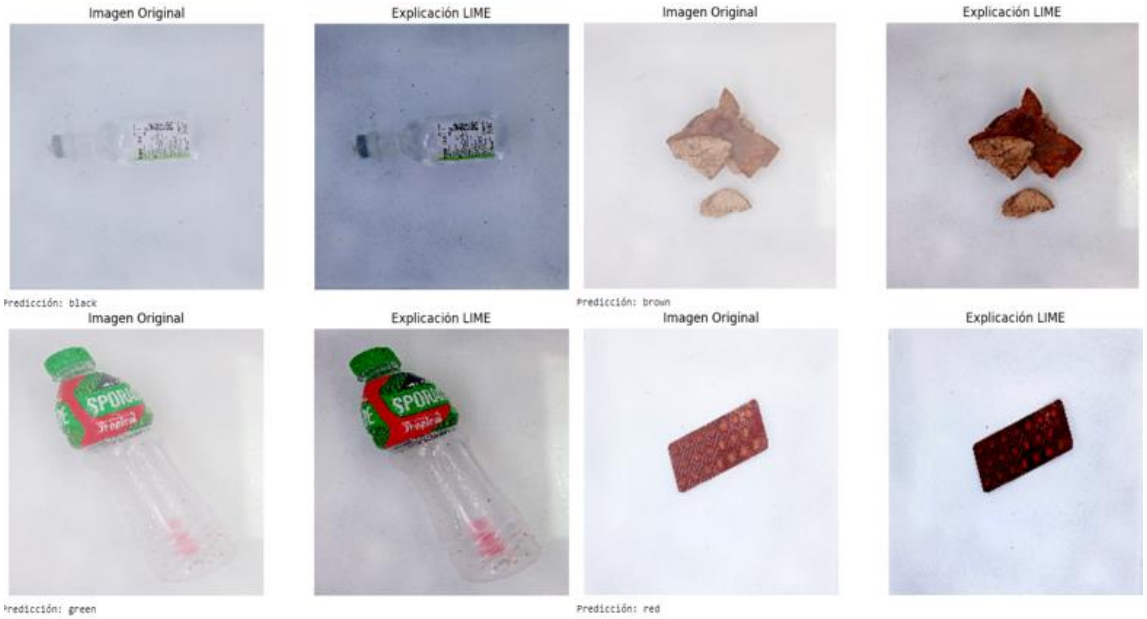


Ilustración 6 Lime aplicado al Dataset de Reusimat_USS_Dataset

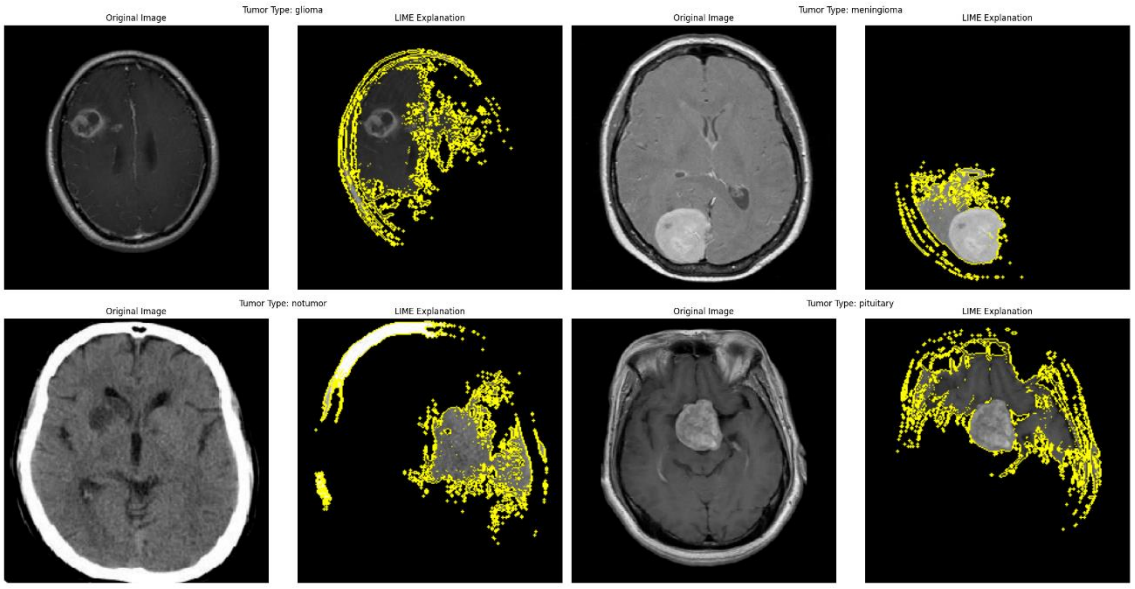


Ilustración 7 Lime aplicado al Dataset de Brain Tumor MRI

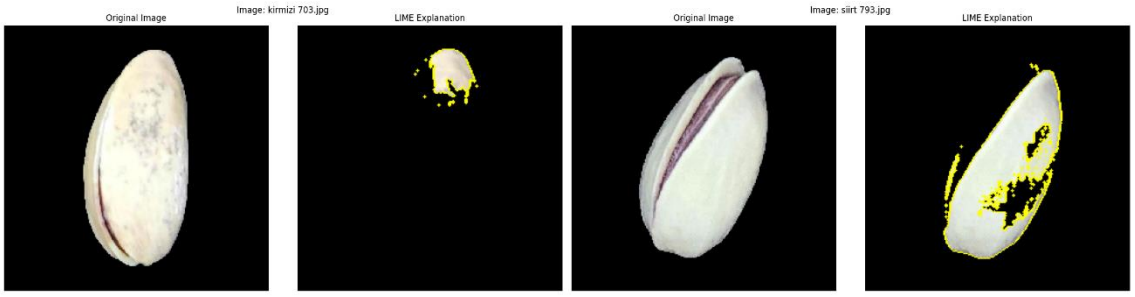


Ilustración 8 Lime aplicado al Dataset de Pistachio

Dataset	Fidelidad	Monotonía	Robustez
Reusimat_USS_Dataset	0.6000	0.0100	0.0200
Brain Tumor MRI	0.4947	1.0000	0.7306
Pistachio	0.4855	1.0000	0.6748

Tabla 11 Resultados de las métricas aplicadas a los datasets con LIME

Método 3: Occlusion Sensitivity

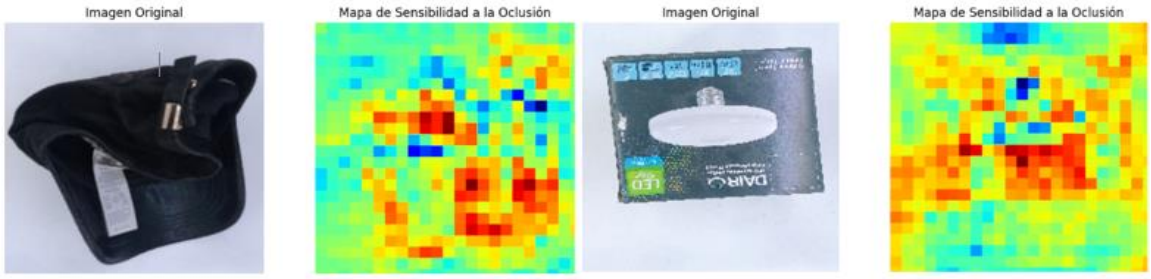


Ilustración 9 Occlusion Sensitivity aplicado al Dataset de Reusimat_USS_Dataset

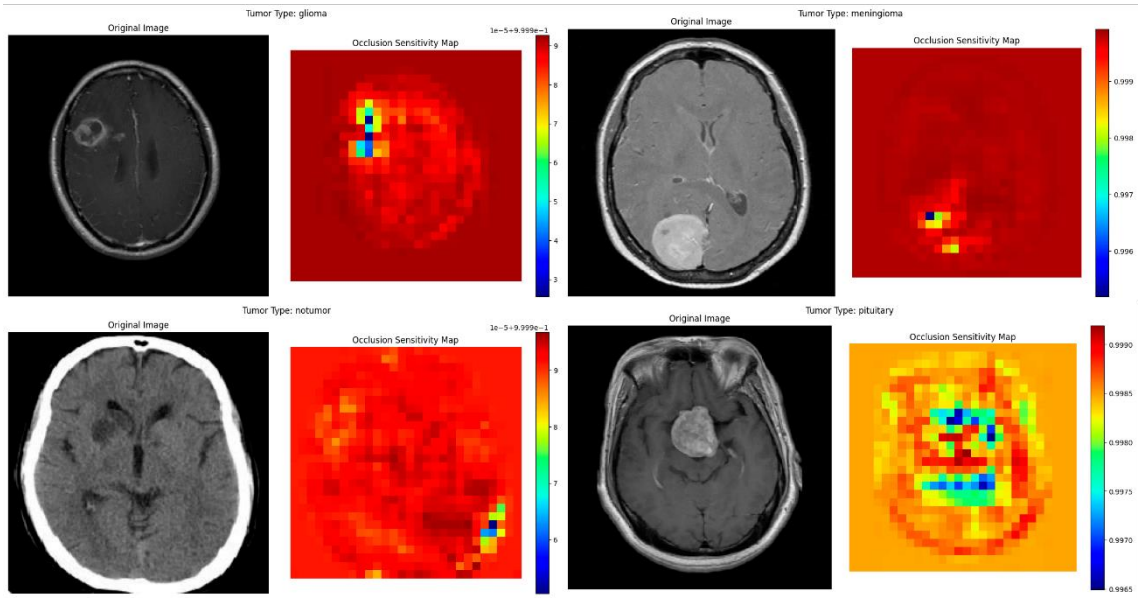


Ilustración 10 Occlusion Sensitivity aplicado al Dataset de Brain Tumor MRI

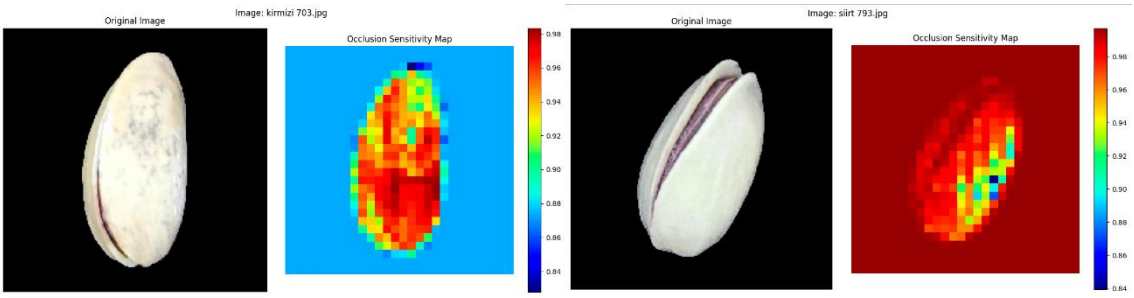


Ilustración 11 Occlusion Sensitivity aplicado al Dataset de Pistachio

Dataset	Fidelidad	Monotonía	Robustez
Reusimat_USS_Dataset	0.2700	0.3200	0.1600
Brain Tumor MRI	0.9419	1.0000	0.9137
Pistachio	0.9400	1.0000	0.8991

Tabla 12 Resultados de las métricas aplicadas a los datasets con Occlusion Sensitivy

IV. DISCUSIÓN Y CONCLUSIONES

4.1. Discusión

Al examinar los resultados obtenidos, se destaca el sólido rendimiento de MobileNetV2 en el dataset Reusimat_USS, alcanzando una accuracy del 95%. Este resultado es particularmente notable considerando la complejidad y diversidad inherentes a las imágenes de residuos sólidos. Por su parte, ResNet50 demostró ser una opción confiable y versátil, logrando accuracies del 97% tanto en el Brain Tumor MRI Dataset como en el Pistachio Dataset, igualando el rendimiento de MobileNetV2 en estos conjuntos de datos. Estos hallazgos resaltan la importancia de considerar las fortalezas específicas de cada modelo en relación con las características particulares de cada dataset.

Adicionalmente, EfficientNetV2B0 exhibió un rendimiento sobresaliente en dos de los datasets evaluados. Este modelo alcanzó una precisión excepcional del 99% tanto en el Brain Tumor MRI Dataset como en el Pistachio Dataset, demostrando una notable eficacia en la clasificación de imágenes médicas y productos agrícolas. Estos resultados subrayan la capacidad de EfficientNetV2B0 para capturar y procesar características complejas en estos dominios específicos, lo que lo posiciona como una opción potente para tareas de clasificación en áreas como la medicina y la agricultura. Es importante destacar que todos los modelos lograron resultados notablemente altos en el dataset de Brain Tumor MRI, lo que indica su potencial para aplicaciones en el campo médico. Esto es particularmente relevante dado el impacto que pueden tener estas herramientas en el diagnóstico y tratamiento de condiciones médicas críticas.

Al examinar las métricas de explicabilidad, observamos patrones que arrojan luz sobre la interpretabilidad de estos modelos. La aplicación de los métodos Grad-CAM, LIME y Occlusion Sensitivity reveló información valiosa sobre cómo estos modelos toman decisiones en diferentes contextos.

En el dataset Reusimat_USS, encontramos una variabilidad significativa en la fidelidad de las explicaciones, con valores que oscilan entre 0.2700 y 0.6000. Esta variabilidad sugiere que la interpretación de los resultados en el contexto de los residuos sólidos es particularmente desafiante, posiblemente debido a la complejidad y diversidad de las imágenes.

Por otro lado, en el Brain Tumor MRI Dataset, observamos una mejora notable en la fidelidad y robustez de las explicaciones, con valores que alcanzaron hasta 0.9419 y 0.9137 respectivamente. Esto indica que los métodos de explicación son particularmente efectivos cuando se aplican a imágenes médicas, proporcionando explicaciones alineadas estrechamente con el proceso de toma de decisiones del modelo.

El Pistachio Dataset mostró resultados similares al de Brain Tumor MRI, con alta fidelidad y robustez en las explicaciones. La consistencia en estos resultados entre dos datasets tan diferentes es alentadora, ya que indica que estos métodos de explicación pueden ser efectivos en una variedad de dominios con características bien definidas.

Un aspecto destacable es la consistencia en la métrica de monotonía, que alcanzó valores de 1.0000 en la mayoría de los casos. Esto sugiere que las explicaciones generadas están perfectamente alineadas con el comportamiento del modelo, proporcionando una representación precisa de cómo el modelo prioriza diferentes características en su proceso de toma de decisiones.

La variabilidad en la robustez de las explicaciones entre los diferentes datasets merece una consideración especial. Mientras que para imágenes médicas y de pistachos las explicaciones parecen ser más robustas, en el caso de los residuos sólidos se observa una mayor variabilidad. Esto podría indicar que la estabilidad de las explicaciones depende en gran medida del dominio específico de las imágenes y de la naturaleza de las características que el modelo está aprendiendo.

Estos hallazgos sobre las métricas de explicabilidad tienen implicaciones importantes para la aplicación práctica de estos modelos. En el campo médico, la alta fidelidad y robustez de las explicaciones podrían proporcionar a los profesionales de la salud información valiosa sobre cómo el modelo interpreta las imágenes de tumores cerebrales. En el contexto de la clasificación de pistachos, las explicaciones confiables podrían ayudar a los agricultores y procesadores a entender mejor los factores que

influyen en la calidad y clasificación de sus productos.

Por otro lado, la variabilidad observada en el dataset Reusimat_USS subraya la necesidad de un enfoque más refinado en la explicabilidad de modelos aplicados a la clasificación de residuos sólidos. Esto podría implicar el desarrollo de métodos de explicación más específicos para este dominio o la exploración de técnicas para mejorar la estabilidad de las explicaciones en conjuntos de datos más heterogéneos.

Este estudio demuestra el potencial significativo de los modelos de aprendizaje profundo en la clasificación de imágenes en diversos dominios, desde la gestión de residuos hasta el diagnóstico médico y el control de calidad agrícola. Los resultados subrayan la importancia de seleccionar cuidadosamente tanto el modelo como el método de explicación según el dominio de aplicación específico. A medida que continuamos mejorando estos modelos y métodos de explicación, nos acercamos a un futuro donde la inteligencia artificial puede proporcionar no solo predicciones precisas, sino también explicaciones interpretables y confiables, lo que podría transformar significativamente diversos sectores de la sociedad.

4.2. Conclusiones

Nuestra investigación sobre métodos de explicación para modelos de aprendizaje profundo en el procesamiento de imágenes digitales ha revelado que no existe un método universalmente superior. Grad-CAM, LIME y Occlusion Sensitivity ofrecen perspectivas complementarias, cada una con sus propias fortalezas en diferentes contextos.

La efectividad de estos métodos varía significativamente según el dominio de aplicación y las características específicas de los datos. Grad-CAM destaca en visualizaciones globales, LIME en interpretaciones locales detalladas, y Occlusion Sensitivity en dominios con características visuales bien definidas.

Esta variabilidad subraya la importancia de un enfoque adaptativo en la selección de métodos de explicación, considerando cuidadosamente el contexto específico y los objetivos de interpretación. La complejidad observada en la interpretabilidad de los modelos refleja la naturaleza intrínseca de los datos que procesan.

Este trabajo no solo contribuye al campo de la explicabilidad en IA, sino que también establece una base sólida para futuras investigaciones que busquen equilibrar el rendimiento de los modelos con su interpretabilidad, avanzando hacia sistemas de IA que sean precisos, transparentes y confiables.

V. REFERENCIAS

[1] C. Pineda Pertuz, “Aprendizaje automático y profundo en Python: una mirada hacia la inteligencia artificial.” RA-MA Editorial.

[2] B. Sistaninejhad, H. Rasi, and P. Nayeri, “A Review Paper about Deep Learning for Medical Image Analysis,” *Computational and Mathematical Methods in Medicine*, vol. 2023. Hindawi Limited, 2023. doi: 10.1155/2023/7091301.

[3] L. Brocki and N. C. Chung, “Feature perturbation augmentation for reliable evaluation of importance estimators in neural networks,” *Pattern Recognit Lett*, vol. 176, 2023, doi: 10.1016/j.patrec.2023.10.012.

[4] D. A. Restrepo Leal, J. P. Viloria Porto, and C. A. Robles Algarín, “El camino a las redes neuronales artificiales,” *El camino a las redes neuronales artificiales*, Sep. 2021, doi: 10.21676/9789587464290.

[5] A. Bosch Rué, J. Casas Roma, and T. Lozano Bagén, *Deep learning: principios y fundamentos*. Editorial UOC, 2019.

[6] C. A. Canelo Sotelo and C. A. Canelo Sotelo, “Redes neuronales artificiales y máquina con soporte vectorial para clasificar a los solicitantes de microcrédito,” *Universidad Nacional de Ingeniería*, 2021, Accessed: May 05, 2024. [Online]. Available: https://repositorio.uni.edu.pe/handle/20.500.14076/22825

[7] A. Oğuz and Ö. F. Ertuğrul, “Introduction to deep learning and diagnosis in medicine,” *Diagnostic Biomedical Signal and Image Processing Applications with Deep Learning Methods*, pp. 1–40, Jan. 2023, doi: 10.1016/B978-0-323-96129-5.00003-2.

[8] L. Nanni, A. Lumini, A. Loreggia, S. Brahnam, and D. Cuza, “Deep ensembles and data augmentation for semantic segmentation,” *Diagnostic Biomedical Signal and Image Processing Applications with Deep Learning Methods*, pp. 215–234, Jan. 2023, doi: 10.1016/B978-0-323-96129-5.00009-3.

[9] B. S. Mackay, K. Marshall, P. Rajendra Kumar, and E. B. K Manash, “Deep learning: a branch of machine learning,” *J Phys Conf Ser*, vol. 1228, no. 1, p. 012045, May 2019, doi: 10.1088/1742-6596/1228/1/012045.

[10] S. Ali *et al.*, “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Information Fusion*, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/J.INFFUS.2023.101805.

[11] P. Hamm, M. Klesel, P. Coberger, and H. F. Wittmann, “Explanation matters: An experimental study on explainable AI,” *Electronic Markets*, vol. 33, no. 1, pp. 1–21, Dec. 2023, doi: 10.1007/S12525-023-00640-9/TABLES/12.

[12] G. Bonifazi *et al.*, “A model-agnostic, network theory-based framework for supporting XAI on classifiers,” *Expert Syst Appl*, vol. 241, p. 122588, May 2024, doi: 10.1016/J.ESWA.2023.122588.

[13] T. Szandała, “Unlocking the black box of CNNs: Visualising the decision-making process with PRISM,” *Inf Sci (N Y)*, vol. 642, p. 119162, Sep. 2023, doi: 10.1016/J.INS.2023.119162.

[14] D. Gragnaniello, F. Marra, L. Verdoliva, and G. Poggi, “Perceptual quality-preserving black-box attack against deep learning image classifiers,” *Pattern Recognit Lett*, vol. 147, pp. 142–149, Jul. 2021, doi: 10.1016/J.PATREC.2021.03.033.

[15] Z. Tian *et al.*, “A Survey of Deep Learning-Based Low-Light Image Enhancement,” *Sensors*, vol. 23, no. 18, Sep. 2023, doi: 10.3390/S23187763.

[16] J. M. Dolezal *et al.*, “Slideflow: deep learning for digital histopathology with real-time whole-slide visualization,” *BMC Bioinformatics*, vol. 25, no. 1, 2024, doi: 10.1186/s12859-024-05758-x.

[17] S. Parvin, S. F. Nimmy, and M. S. Kamal, “Convolutional neural network based data interpretable framework for Alzheimer’s treatment planning,” *Vis Comput Ind Biomed*

Art, vol. 7, no. 1, 2024, doi: 10.1186/s42492-024-00154-x.
[18] F. Qayyum, N. A. Samee, M. Alabdulhafith, A. Aziz, and M. Hijjawi, “Shapley-based interpretation of deep learning models for wildfire spread rate prediction,” *Fire Ecology*, vol. 20, no. 1, 2024, doi: 10.1186/s42408-023-00242-y.
[19] P. Borole and A. Rajan, “Building trust in deep learning-based immune response predictors with interpretable explanations,” *Commun Biol*, vol. 7, no. 1, 2024, doi: 10.1038/s42003-024-05968-2.
[20] S. Prabhu, K. Prasad, T. Hoang, X. Lu, and S. I., “Multi-organ squamous cell carcinoma classification using feature interpretation technique for explainability,” *Biocybern Biomed Eng*, vol. 44, no. 2, pp. 312–326, 2024, doi: 10.1016/j.bbe.2024.03.001.
[21] R. Romero-Oraá, M. Herrero-Tudela, M. I. López, R. Hornero, and M. García, “Attention-based deep learning framework for automatic fundus image processing to aid in diabetic retinopathy grading,” *Comput Methods Programs Biomed*, vol. 249, 2024, doi: 10.1016/j.cmpb.2024.108160.
[22] C. Metta *et al.*, “Advancing Dermatological Diagnostics: Interpretable AI for Enhanced Skin Lesion Classification,” *Diagnostics*, vol. 14, no. 7, 2024, doi: 10.3390/diagnostics14070753.
[23] R. Haque, M. M. Hassan, A. K. Bairagi, and S. M. Shariful Islam, “NeuroNet19: an explainable deep neural network model for the classification of brain tumors using magnetic resonance imaging data,” *Sci Rep*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-51867-1.
[24] F. Ahmed *et al.*, “Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence,” *Sci Rep*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-56478-4.
[25] T. Chanda *et al.*, “Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma,” *Nat Commun*, vol. 15, no. 1, 2024, doi: 10.1038/s41467-023-43095-4.
[26] E. Agbozo and D. M. Balungu, “Liver Disease Classification - An XAI Approach to Biomedical AI,” *Informatica (Slovenia)*, vol. 48, no. 1, pp. 79–90, 2024, doi: 10.31449/inf.v48i1.4611.
[27] T. Iqbal, A. Khalid, and I. Ullah, “Explaining decisions of a light-weight deep neural network for real-time coronary artery disease classification in magnetic resonance imaging,” *J Real Time Image Process*, vol. 21, no. 2, 2024, doi: 10.1007/s11554-023-01411-7.
[28] Z. J. Lo *et al.*, “Development of an explainable artificial intelligence model for Asian vascular wound images,” *Int Wound J*, vol. 21, no. 4, 2024, doi: 10.1111/iwj.14565.
[29] R. P. Ethiraj and K. Paranjothi, “A deep learning-based approach for early detection of disease in sugarcane plants: an explainable artificial intelligence model,” *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 974–983, 2024, doi: 10.11591/ijai.v13.i1.pp974-983.
[30] J. Domínguez *et al.*, “ROAD2H: Development and evaluation of an open-source explainable artificial intelligence approach for managing co-morbidity and clinical guidelines,” *Learn Health Syst*, vol. 8, no. 2, 2024, doi: 10.1002/lrh2.10391.
[31] A. Kodipalli, S. L. Fernandes, and S. Dasar, “An Empirical Evaluation of a Novel Ensemble Deep Neural Network Model and Explainable AI for Accurate Segmentation and Classification of Ovarian Tumors Using CT Images,” *Diagnostics*, vol. 14, no. 5, 2024, doi: 10.3390/diagnostics14050543.
[32] E. Cerekci *et al.*, “Quantitative evaluation of Saliency-Based Explainable artificial intelligence (XAI) methods in Deep Learning-Based mammogram analysis,” *Eur J Radiol*, vol. 173, 2024, doi: 10.1016/j.ejrad.2024.111356.
[33] J. Gu, Y. Yang, and V. Tresp, *Understanding Individual Decisions of CNNs via Contrastive Backpropagation*, vol. 11363 LNCS. 2019. doi: 10.1007/978-3-030-20893-6_8.

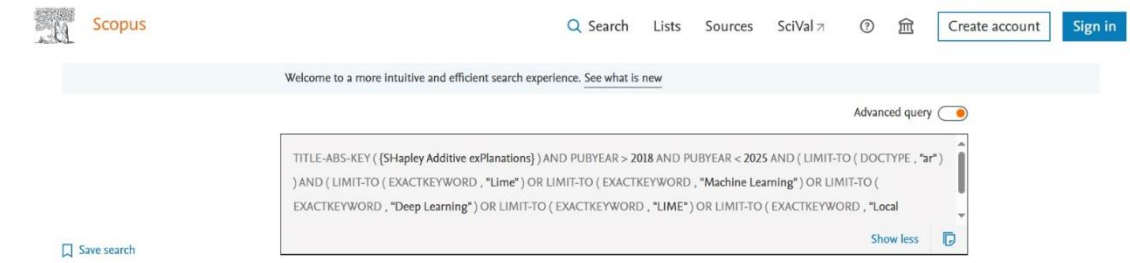
[34] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso, “Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer’s Disease,” in *Proceedings of the International Joint Conference on Neural Networks*, 2020. doi: 10.1109/IJCNN48605.2020.9206837.
[35] V. Swamy, S. Du, M. Marras, and T. Kaser, “Trusting the Explainers: Teacher Validation of Explainable Artificial Intelligence for Course Design,” in *ACM International Conference Proceeding Series*, 2023, pp. 345–356. doi: 10.1145/3576050.3576147.
[36] S. Bengamra, E. Zagrouba, and A. Bigand, “Explainable AI for Deep Learning Based Potato Leaf Disease Detection,” in *IEEE International Conference on Fuzzy Systems*, 2023. doi: 10.1109/FUZZ52849.2023.10309803.
[37] F. G. Ringwald, A. Martynova, J. Mierisch, M. Wielpütz, and U. Eisenmann, *Explainable Artificial Intelligence for Deep-Learning Based Classification of Cystic Fibrosis Lung Changes in MRI*, vol. 310. 2024. doi: 10.3233/SHTI231099.
[38] K. Letrache and M. Ramdani, “Explainable Artificial Intelligence: A Review and Case Study on Model-Agnostic Methods,” in *Proceedings - SITA 2023: 2023 14th International Conference on Intelligent Systems: Theories and Applications*, 2023. doi: 10.1109/SITA60746.2023.10373722.
[39] L. Holmberg, “‘When Can I Trust It?’ Contextualising Explainability Methods for Classifiers,” in *ACM International Conference Proceeding Series*, 2023, pp. 108–115. doi: 10.1145/3589883.3589899.
[40] Y. Huang, N. Schaal, M. Hefenbrock, Y. Zhou, T. Riedel, and M. Beigl, “McXai: Local Model-Agnostic Explanation As Two Games,” in *Proceedings of the International Joint Conference on Neural Networks*, 2023. doi: 10.1109/IJCNN54540.2023.10191756.
[41] K. Dawoud, W. Samek, P. Eisert, S. Lapuschkin, and S. Bosse, “Human-Centered Evaluation of XAI Methods,” in *IEEE International Conference on Data Mining Workshops, ICDMW*, 2023, pp. 912–921. doi: 10.1109/ICDMW60847.2023.00122.
[42] B. Finzel, I. Rieger, S. Kuhn, and U. Schmid, *Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition*, vol. 14065 LNCS. 2023. doi: 10.1007/978-3-031-40837-3_3.
[43] “Brain Tumor MRI Dataset.” Accessed: Jul. 02, 2024. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data
[44] “Pistachio Image Dataset.” Accessed: Jul. 02, 2024. [Online]. Available: https://www.kaggle.com/datasets/muratkokludataset/pistachio-image-dataset
[45] Y. Wang *et al.*, “A Gradient Mapping Guided Explainable Deep Neural Network for Extracapsular Extension Identification in 3D Head and Neck Cancer Computed Tomography Images,” Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.00895
[46] K. Lamba and S. Rani, “A novel approach of brain-computer interfacing (BCI) and Grad-CAM based explainable artificial intelligence: Use case scenario for smart healthcare,” *J Neurosci Methods*, vol. 408, Aug. 2024, doi: 10.1016/j.jneumeth.2024.110159.
[47] M. M. M, M. T. R, V. K. V, and S. Guluwadi, “Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50,” *BMC Med Imaging*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12880-024-01292-7.
[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” Oct. 2016, doi: 10.1007/s11263-019-01228-7.
[49] M. E. Klontzas, G. Kalarakis, E. Koltsakis, T. Papathomas, A. H. Karantanas, and A. Tzortzakakis, “Convolutional neural networks for the differentiation between benign and malignant renal tumors with a multicenter international computed tomography dataset,” *Insights Imaging*, vol. 15, no. 1, Dec. 2024, doi: 10.1186/s13244-023-01601-8.
[50] M. S. Ayhan, J. Neubauer, M. M. Uzel, F. Gelisken, and P. Berens, “Interpretable detection of epiretinal membrane from optical coherence tomography with deep neural networks,” *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-57798-1.
[51] Y. Yan, T. Jiang, X. Li, L. Sun, J. Zhu, and J. Lin, “Model-agnostic progressive saliency map

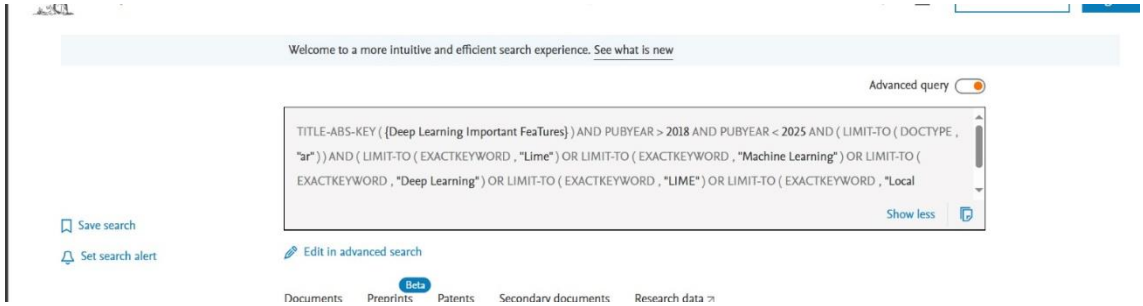
generation for object detector," *Image Vis Comput*, vol. 145, May 2024, doi: 10.1016/j.imavis.2024.104988.
[52] B. U. Maheswari *et al.*, "Explainable deep-neural-network supported scheme for tuberculosis detection from chest radiographs," *BMC Med Imaging*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12880-024-01202-x.
[53] S. Zhou *et al.*, "Interpretable machine learning model for early prediction of 28-day mortality in ICU patients with sepsis-induced coagulopathy: development and validation," *Eur J Med Res*, vol. 29, no. 1, Dec. 2024, doi: 10.1186/s40001-023-01593-7.
[54] S. K. Ghosh and A. H. Khandoker, "Investigation on explainable machine learning models to predict chronic kidney diseases," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-54375-4.
[55] C. C. Chang, T. C. Liu, C. J. Lu, H. C. Chiu, and W. N. Lin, "Explainable machine learning model for identifying key gut microbes and metabolites biomarkers associated with myasthenia gravis," *Comput Struct Biotechnol J*, vol. 23, pp. 1572–1583, Dec. 2024, doi: 10.1016/j.csbj.2024.04.025.
[56] A. A. Abujaber, Y. Imam, I. Albalkhi, S. Yaseen, A. J. Nashwan, and N. Akhtar, "Utilizing machine learning to facilitate the early diagnosis of posterior circulation stroke," *BMC Neurol*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12883-024-03638-8.
[57] J. Li, Y. Zhai, Y. Cao, Y. Xia, and R. Yu, "Development of an interpretable machine learning model associated with genetic indicators to identify Yin-deficiency constitution," *Chin Med*, vol. 19, no. 1, May 2024, doi: 10.1186/s13020-024-00941-x.
[58] J. Kim, J. Lee, J. Seo, Y. Lee, and Y. S. Na, "Disruption prediction and analysis through multimodal deep learning in KSTAR," *Fusion Engineering and Design*, vol. 200, Mar. 2024, doi: 10.1016/j.fusengdes.2024.114204.
[59] Q. Ouyang *et al.*, "Machine learning-coupled tactile recognition with high spatiotemporal resolution based on cross-striped nanocarbon piezoresistive sensor array," *Biosens Bioelectron*, vol. 246, Feb. 2024, doi: 10.1016/j.bios.2023.115873.
[60] L. Chen, Y. Zhao, J. Qiu, and X. Lin, "Analysis and validation of biomarkers of immune cell-related genes in postmenopausal osteoporosis: An observational study," *Medicine*, vol. 103, no. 19, p. e38042, May 2024, doi: 10.1097/MD.0000000000038042.
[61] J. H. Shin, J. Bae, J. M. Kim, and S. J. Lee, "An interpretable convolutional neural network for nuclear power plant abnormal events," *Appl Soft Comput*, vol. 132, Jan. 2023, doi: 10.1016/j.asoc.2022.109792.
[62] H. J. Cho, M. Shu, S. Bekiranov, C. Zang, and A. Zhang, "Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment," *Bioinformatics*, vol. 39, no. 4, Apr. 2023, doi: 10.1093/bioinformatics/btad113.
[63] J. Hu *et al.*, "Machine learning models to predict electroencephalographic seizures in critically ill children," *Seizure*, vol. 87, pp. 61–68, Apr. 2021, doi: 10.1016/j.seizure.2021.03.001.
[64] "AI Explainability 360." Accessed: Jul. 13, 2024. [Online]. Available: https://aix360.res.ibm.com/

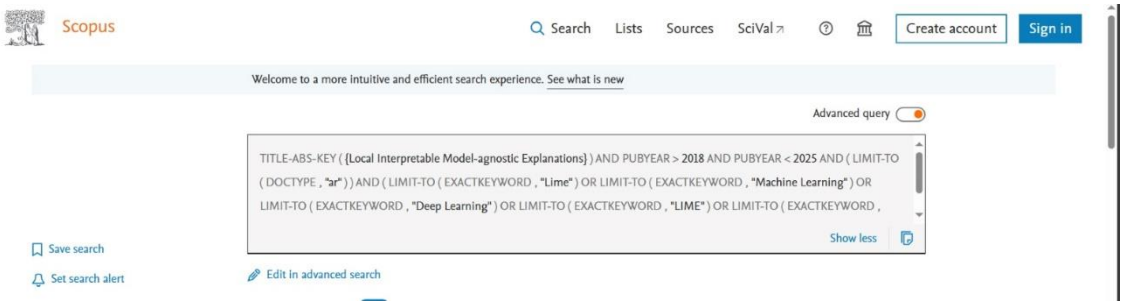
ANEXOS

ANEXO 01: INSTRUMENTOS DE RECOLECCIÓN DE DATOS

✓ Estrategia de búsqueda







✓ Plantilla para la síntesis

Artículo	Método	Ref	Resultados	Descripción
A Gradient Mapping Guided Explainable Deep Neural Network for Extracapsular Extension Identification in 3D Head and Neck Cancer Computed Tomography Images	Gradient Mapping Guided Explainable Network (GMGENet)	[1, 2, 3, 4, 5, 6]	Precisión: 0.902 (90.2%) [6] AUC: 0.911 (91.1%) [6] Sensibilidad: 0.909 (90.9%) [6] Especificidad: 0.895 (89.5%) [6]	GMGENet utiliza Grad-CAM para guiar la red neuronal profunda a concentrarse en las regiones relevantes para la identificación de extensión extracapsular (ECE) en imágenes de CT 3D, extrayendo volúmenes de interés (VOIs) sin anotación manual.
A novel approach of brain-computer interfacing (BCI) and Grad-CAM based explainable artificial intelligence: Use case scenario for smart healthcare	Xception con Grad-CAM para BCI	[1, 2, 3, 4, 5, 6]	Precisión: 98.92% [6] Sensibilidad: 99.09% [6] Especificidad: 98.18% [6] F1 Score: 98.91% [6]	Xception emplea transfer learning sobre el dataset de MRI cerebral, utilizando Grad-CAM para generar heatmaps que resaltan las características responsables de las predicciones. El modelo mejora la transparencia en las decisiones de la BCI.
Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50	ResNet50 con Grad-CAM	[22, 23, 24, 25]	Precisión: 98.52% [25] Sensibilidad: 97.8% [25] Especificidad: 99.1% [25] AUC: 0.993 [25] F1 Score: 0.981 [25]	Utiliza la arquitectura ResNet50 para detectar tumores cerebrales en imágenes de MRI. Grad-CAM se emplea para generar mapas de activación que destacan las regiones de interés en las imágenes, proporcionando interpretabilidad al modelo. La combinación de ResNet50 y Grad-CAM mejora la detección de tumores cerebrales mediante la integración de técnicas avanzadas de aprendizaje profundo y visualización.
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization	Grad-CAM (Gradient-weighted Class Activation Mapping)	[59, 53, 52, 51, 33	Precisión en localización: 56.51% (VGG-16) [59] AUC en segmentación débil: 0.496 [59] Correlación con mapas de oclusión: 0.261 [59]	Grad-CAM utiliza gradientes del concepto objetivo que fluyen en la última capa convolucional para producir un mapa de localización que resalta las regiones importantes en la imagen para predecir el concepto. Se combina con Guided Grad-CAM para crear visualizaciones de alta resolución y discriminativas de clase, aplicables a varios modelos CNN sin necesidad de cambios arquitectónicos o reentrenamiento.

Artículo	Método	Ref	Resultados	Descripción
Explainable deep-neural-network supported scheme for tuberculosis detection from chest radiographs	Shallow-CNN con CAM y LIME	[23, 24, 6, 13, 22]	Precisión: 95% [13] F1 Score: 0.95 [13] AUC: 0.976 [13] Sensibilidad: 0.95 [13] Especificidad: 0.95 [13]	Se desarrolla un Shallow-CNN con cuatro capas de convolución y pooling, optimizado mediante la técnica de optimización bayesiana, para la detección de tuberculosis en radiografías de tórax. Utiliza Class Activation Maps (CAM) y Local Interpretable Model-agnostic Explanations (LIME) para mejorar la interpretabilidad del modelo, mostrando las áreas de interés en las imágenes que contribuyen a la clasificación.
Interpretable machine learning model for early prediction of 28-day mortality in ICU patients with sepsis-induced coagulopathy: development and validation	XGBoost con SHAP y LIME	[10, 14, 15, 16, 17]	AUROC: 0.828, 0.913, 0.923 [10] AUPRC: 0.807, 0.796, 0.921 [10] Precisión: 0.785, 0.885, 0.891 [10] F1 Score: 0.63, 0.69, 0.70 [10]	XGBoost se utiliza para predecir la mortalidad a 28 días en pacientes con coagulopatía inducida por sepsis (SIC) en las bases de datos MIMIC-III, MIMIC-IV y eICU-CRD. SHAP y LIME se aplican para proporcionar interpretaciones de las predicciones, destacando la importancia de características como la puntuación SOFA, RDW, y edad en el modelo.
Investigation on explainable machine learning models to predict chronic kidney diseases	XGBoost con SHAP y LIME	[31, 32, 33, 36, 37]	Precisión: 93.29% [37] Sensibilidad: 91.80% [37] Especificidad: 94.73% [37] AUC: 0.9689 [37] F1 Score: 0.9313 [37]	XGBoost se utiliza para predecir la enfermedad renal crónica (CKD) utilizando características clínicas y de laboratorio. SHAP y LIME se emplean para explicar la influencia de las características en las predicciones del modelo, proporcionando una comprensión detallada de cómo las características individuales afectan los resultados del modelo.

Artículo	Método	Ref	Resultados	Descripción
Convolutional neural networks for the differentiation between benign and malignant renal tumors with a multicenter international computed tomography dataset	Convolutional Neural Networks (InceptionV3, Inception-ResNetV2, VGG-16) con Grad-CAM	[23, 24, 25, 26, 27]	Inception-ResNetV2 Precisión: 95.18% [27] Sensibilidad: 90.35% [27] Especificidad: 100% [27] AUC: 0.918 (95% CI 0.873–0.963) [27] F1 Score: 96.6% [27]	Se utilizan las arquitecturas InceptionV3, Inception-ResNetV2 y VGG-16 preentrenadas con ImageNet y ajustadas con el dataset específico para diferenciar entre tumores renales benignos y malignos. Grad-CAM se emplea para crear mapas de saliencia que destacan las características relevantes en las imágenes CT. El modelo Inception-ResNetV2 mostró el mejor desempeño, enfocándose en la interfaz entre el tumor y el parénquima renal circundante, mejorando así la precisión en la clasificación.
Interpretable detection of epiretinal membrane from optical coherence tomography with deep neural networks	Ensemble de Redes Neuronales Profundas (DNNs) con Saliency Maps	[19, 20, 22, 25, 27]	AUC para no-ERM: 0.99 [27] AUC para pequeño-ERM: 0.92 [27] AUC para grande-ERM: 0.99 [27] Precisión 3-way: 89.33% [27] Sensibilidad: 95.45% [27]	El estudio utiliza un ensamble de redes neuronales profundas (DNNs) con las arquitecturas ResNet50 e InceptionV3 para detectar y clasificar membranas epirretinales (ERM) en imágenes de OCT. Los mapas de saliencia generados con Guided-Backprop resaltan áreas importantes en las imágenes OCT para una interpretación clara de las decisiones del modelo.
Model-agnostic progressive saliency map generation for object detector	Generación Progresiva de Mapas de Saliencia Model-Agnostic (MAPSM)	[27, 28, 29, 30, 31]	Deletion Score: 0.045 [31] Insertion Score: 0.818 [31] Mean Saliency: 0.085 [31] Saliency Average Contribution (SAC): 2.39 [31]	MAPSM es un método de generación de mapas de saliencia basado en un marco jerárquico para modelos de detección de objetos. A diferencia de otros métodos de caja negra, MAPSM introduce una partición adaptativa de máscaras y una estrategia de generación de máscaras impulsada por la saliencia para reducir el ruido en los mapas de saliencia. Progresivamente descubre y refina las áreas de saliencia de los objetos, resultando en mapas de saliencia más interpretables y de mejor calidad.

Artículo	Método	Ref	Resultados	Descripción
Explainable machine learning model for identifying key gut microbes and metabolites biomarkers associated with myasthenia gravis	Random Forest (RF) con SHAP	[29, 30, 31, 32, 33]	Precisión: 89.66% [33] Sensibilidad: 89.47% [33] Especificidad: 90.00% [33] AUC: 0.9421 [33] F1 Score: 91.89% [33]	Se utiliza el algoritmo Random Forest para construir modelos de predicción utilizando datos de microbiota intestinal y metabolitos. SHAP se aplica para proporcionar explicaciones interpretables a nivel global y personalizado, permitiendo entender la influencia de características individuales en las predicciones del modelo.
Utilizing machine learning to facilitate the early diagnosis of posterior circulation stroke	XGBoost con SHAP	[33, 34, 35, 36, 37]	AUC: 0.81 [37] Precisión: 0.79 [37] Sensibilidad: 0.62 [37] Especificidad: 0.83 [37] F1 Score: 0.55 [37]	Se utiliza XGBoost para predecir la probabilidad de diagnóstico de síndrome de circulación posterior (PCS) utilizando datos clínicos. SHAP se emplea para interpretar las predicciones del modelo, identificando las características más influyentes en la clasificación de PCS, como el IMC, glucosa en sangre, ataxia, disartria, presión arterial diastólica y temperatura corporal.
Development of an interpretable machine learning model associated with genetic indicators to identify Yin-deficiency constitution	Random Forest con SHAP	[23, 24, 25, 26, 27]	AUC: 0.937 (95% CI 0.844–1.000) [26] Sensibilidad: 0.870 [26] Especificidad: 0.900 [26	Se utiliza el algoritmo Random Forest para construir modelos predictivos utilizando los indicadores genéticos NFKBIA, BCL2A1, y CCL4 para la identificación de la constitución Yin-deficiencia (YinDC). SHAP se aplica para proporcionar explicaciones interpretables, permitiendo visualizar la contribución de cada predictor a las predicciones del modelo y facilitando la comprensión de los resultados del modelo.

Artículo	Método	Ref	Resultados	Descripción
Disruption prediction and analysis through multimodal deep learning in KSTAR	Multimodal Deep Learning con ViViT y Transformer	[16, 17, 18, 19, 22]	Precisión: 88.7% [22] AUC: 0.91 [22] F1 Score: 0.89 [22] Sensibilidad: 87.4% [22] Especificidad: 90.2% [22]	Se utiliza un enfoque de aprendizaje profundo multimodal que combina datos de video de IVIS y parámetros 0D de plasma para predecir las interrupciones en KSTAR. ViViT maneja los datos de video y Transformer maneja los datos 0D. GradCAM y Attention Rollout se utilizan para evaluar la capacidad de los modelos para capturar características relevantes para la predicción de interrupciones.
Machine learning-coupled tactile recognition with high spatiotemporal resolution based on cross-striped nanocarbon piezoresistive sensor array	Red Neuronal con TSNE y Red Neuronal Unicapa	[24, 25, 26, 27, 28]	Precisión: 98.9% [28] Discernibilidad: 98.9% [28] AUC: >0.999 [28] Temporal resolution: 4.5 ms [26] Sensibilidad: 11.4 (mv/Kpa) [26]	Se desarrolló una red neuronal combinada con TSNE para visualizar la diferencia entre imágenes de distintas formas en la matriz sensora de presión piezorresistiva (PRSA) y una red neuronal unicapa para cuantificar la discernibilidad entre las imágenes de diferentes formas. El sistema de sensores muestra una alta resolución espaciotemporal y excelente desempeño en la visualización en tiempo real de múltiples puntos de contacto y seguimiento de la trayectoria del movimiento.
Analysis and validation of biomarkers of immune cell-related genes in postmenopausal osteoporosis: An observational study	Random Forest (RF) con LASSO y qRT-PCR	[19, 20, 21, 22, 23]	Precisión: 93.50% [23] Sensibilidad: 94.12% [23] Especificidad: 92.85% [23] AUC: 0.967 (95% CI 0.930–0.990) [23]	Utiliza Random Forest combinado con LASSO para identificar genes biomarcadores clave en la osteoporosis postmenopáusica (PMOP). La validación experimental se realiza con qRT-PCR en muestras de sangre de pacientes con PMOP y controles sanos. Este enfoque permite identificar genes clave relacionados con células inmunes que son discriminativos entre altos y bajos niveles de densidad mineral ósea.

Artículo	Método	Ref	Resultados	Descripción
An interpretable convolutional neural network for nuclear power plant abnormal events	CNN con Saliency Mapping, Guided Grad-CAM y DeepSHAP	[10, 11, 12, 13, 16]	Precisión: 99.33% [16] Sensibilidad: 98.76% [16] Especificidad: 100% [16] AUC: 0.997 [16]	Se utiliza una red neuronal convolucional (CNN) para diagnosticar 10 estados anormales en plantas nucleares. Para aumentar la transparencia del modelo, se aplican tres técnicas explicativas: saliency mapping, Guided Grad-CAM, y DeepSHAP. Estas técnicas identifican los parámetros de entrada más influyentes en la clasificación, optimizando la cantidad de parámetros a monitorear para el diagnóstico eficiente de estados anormales
Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment	Meta-Learning con DeepLIFT	[1, 2, 3, 5, 10]	C-Index (Individual Datasets): 0.74 (Transcriptómica) [3], 0.75 (Clínica) [3], 0.58 (Proteómica) [3] C-Index (Combinaciones): 0.79 (Integrado) [3], 0.84 (Clínica + Transcriptómica) [3], 0.63 (Proteómica + Transcriptómica) [3] Integrated Brier Score: 0.12 [3] Enriquecimiento de vías: DNA repair pathways (P < 0.05) [5]	Utiliza un modelo de meta-aprendizaje combinado con DeepLIFT para el análisis de supervivencia basado en datos multi-ómicos de cáncer. El modelo mejora el análisis de supervivencia al integrar transcriptómica, proteómica y datos clínicos de TCGA, con interpretabilidad proporcionada por DeepLIFT, que asigna puntuaciones de contribución a los genes para identificar vías moleculares relevantes.
Machine learning models to predict electroencephalographic seizures in critically ill children	Random Forest, LASSO, y DeepLIFT	[54, 55, 56, 57]	Precisión: 0.785 (Random Forest) [56] AUROC: 0.781 (LASSO) [56] F1 Score: 0.758 (DeepLIFT) [56] Validación Cruzada: 0.792 (Promedio) [56]	Se utilizan Random Forest, LASSO y DeepLIFT para predecir la ocurrencia de convulsiones electroencefalográficas (ES) en niños críticamente enfermos. Random Forest construye un modelo a partir de múltiples árboles de decisión, LASSO aplica regularización para seleccionar variables clave, y DeepLIFT asigna puntuaciones de importancia a las características basadas en redes neuronales, permitiendo interpretar la contribución de cada característica en la predicción.