ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.1355

EXPLAINABLE NEURAL NETWORKS: TRANSPARENCY AND TRUST IN MEDICAL DIAGNOSIS WITH RADIOLOGICAL IMAGES: A SYSTEMATIC REVIEW

Miguel Hans Paiva Sanchez[1], Eduardo Aldair, Mendoza Crisanto[2], Universidad Tecnológica del Perú, Perú, U20223496@utp.edu.pe, U20212086@utp.edu.pe

***Abstract*- Advances in Explainable Artificial Intelligence (XAI) have reshaped the landscape of medical diagnosis by addressing challenges related to the interpretability and trust of deep learning models. This study conducts a Systematic Literature Review (SLR), following PRISMA guidelines and the PICO strategy, to examine how XAI techniques have been applied to the analysis of radiological images, including X-rays, computed tomography (CT), and magnetic resonance imaging (MRI). The review involved searches across Scopus, Redalyc, and SciELO databases, initially identifying 230 articles, of which 40 met the inclusion criteria.**

The analysis revealed a predominance of post-hoc interpretability techniques such as SHAP and LIME, as well as model-inherent approaches including neural networks with attention mechanisms. These methods have demonstrated significant improvements in diagnostic precision and have enhanced clinical interpretability. However, the review also identified critical gaps in the standardization of evaluation metrics and the usability of generated explanations in clinical practice. This work provides key recommendations to guide future research toward the development of explainability tools tailored to the needs of healthcare professionals, fostering greater trust in AI-supported medical diagnosis.

Keywords-- Medical Diagnosis, Radiological Imaging, Explainable Artificial Intelligence.

REDES NEURONALES EXPLICABLES: TRANSPARENCIA Y CONFIANZA EN EL DIAGNÓSTICO MÉDICO CON IMÁGENES RADIOLÓGICAS: UNA REVISIÓN SISTEMÁTICA

Miguel Hans Paiva Sanchez[1], Eduardo Aldair, Mendoza Crisanto[2], Universidad Tecnológica del Perú, Perú, U20223496@utp.edu.pe, U20212086@utp.edu.pe

Resumen- ***Los avances en la inteligencia artificial explicable (XAI, por sus siglas en inglés) han transformado el ámbito del diagnóstico médico, abordando los retos de interpretabilidad y confianza asociados con los modelos de aprendizaje profundo. Este estudio realiza una revisión sistemática de la literatura (RSL), basada en las directrices PRISMA y la estrategia PICO, con el objetivo de analizar cómo la XAI ha sido aplicada en el procesamiento e interpretación de imágenes radiológicas, tales como radiografías, tomografías computarizadas y resonancias magnéticas. La búsqueda se realizó en las bases de datos Scopus, Redalyc y SciELO, identificándose inicialmente 230 estudios, de los cuales 40 cumplieron con todos los criterios de elegibilidad.***

El análisis de contenido evidenció un predominio de técnicas post-hoc, como SHAP y LIME, así como enfoques inherentes al modelo, como redes neuronales con mecanismos de atención. Estos métodos han demostrado mejoras sustanciales en la precisión diagnóstica y en la interpretación clínica de los resultados. Sin embargo, se identificaron brechas importantes en la estandarización de métricas de evaluación y en la adecuación de las explicaciones generadas, limitando su adopción efectiva en entornos hospitalarios. La presente revisión ofrece recomendaciones clave orientadas al desarrollo de herramientas explicables centradas en las necesidades del personal médico, fortaleciendo así la confianza en la inteligencia artificial aplicada al diagnóstico médico

Palabras claves—Diagnóstico Médico, Imágenes Radiológicas, Inteligencia Artificial Explicable, IA.

I. INTRODUCCIÓN

El uso de la inteligencia artificial (IA) en la medicina ha experimentado un crecimiento significativo, especialmente en la interpretación de imágenes radiológicas. Las redes neuronales profundas han permitido mejorar la precisión en el diagnóstico de enfermedades como el cáncer, afecciones pulmonares y patologías cardíacas [1], [2], [4]. Estas tecnologías han logrado incluso superar el desempeño humano en tareas específicas, optimizando el tiempo de diagnóstico y la toma de decisiones clínicas [6], [9], [14]. Sin embargo, una de las principales barreras para su adopción en entornos hospitalarios es la falta de interpretabilidad de sus modelos. Al operar como "cajas negras", estos sistemas generan desconfianza entre los profesionales de la salud, quienes no comprenden completamente cómo se generan los resultados [3], [7].

En este contexto, la inteligencia artificial explicable (XAI, por sus siglas en inglés) ha emergido como una solución fundamental para mitigar esta problemática. Su objetivo es dotar a los modelos de aprendizaje profundo de capacidades explicativas que permitan interpretar sus decisiones de manera comprensible para los médicos [10], [15], [23]. En los últimos años, se han propuesto técnicas como SHAP, LIME y redes neuronales con mecanismos de atención, que ofrecen interpretaciones visuales o estructuradas de los diagnósticos generados por IA [17], [22], [29], [36], [40]. Estas herramientas no solo buscan mejorar la transparencia de los modelos, sino también reforzar la confianza del personal clínico en la tecnología.

No obstante, persisten vacíos importantes en la literatura científica sobre cómo integrar la explicabilidad sin comprometer la precisión diagnóstica. Existen pocos estudios que analicen de forma sistemática las diferentes metodologías de XAI aplicadas específicamente al análisis de imágenes radiológicas, lo cual limita su adopción clínica a gran escala [16], [18], [34].

Frente a esta brecha, el presente estudio desarrolla una Revisión Sistemática de Literatura (RSL), siguiendo la metodología PRISMA y la estrategia de búsqueda PICO, con el objetivo de identificar, analizar y comparar las principales aplicaciones de XAI en radiografías, tomografías computarizadas y resonancias magnéticas. Se analizaron 40 artículos seleccionados entre más de 200 recuperados inicialmente desde bases como Scopus, SciELO y Redalyc.

Este artículo se organiza de la siguiente manera: la Sección II describe la metodología utilizada, incluyendo los criterios de inclusión, exclusión y el proceso de selección. La Sección III presenta los resultados, segmentados en análisis bibliométrico y temático. En la Sección IV, se discuten los principales hallazgos, destacando avances, desafíos y oportunidades de mejora. Finalmente, la Sección V ofrece las conclusiones y recomendaciones orientadas al desarrollo de soluciones explicables y confiables en el contexto clínico.

II. METODOLOGÍA

Con la finalidad de analizar cómo la inteligencia artificial (IA) mejora el diagnóstico médico basado en imágenes radiológicas mediante el uso de redes neuronales explicables, se desarrolló una búsqueda de revisión sistemática. Esta revisión tiene como objetivo organizar y estructurar la búsqueda de manera rigurosa, empleando la estrategia PICO, que facilita la definición clara de los pasos y acciones necesarios para abordar la pregunta de investigación [4].

La pregunta PICO se centra en evaluar el impacto de las redes neuronales explicables en la mejora de la transparencia y confianza en el diagnóstico médico, en comparación con los métodos tradicionales. En la Tabla N° 1, se presenta la generación de la pregunta PICO junto a los componentes y subpreguntas que orientan el análisis.

La metodología PICO permitió estructurar la revisión mediante la definición de cuatro componentes: población (P), intervención (I), comparación (C) y resultado (O). Esto facilitó la formulación de subpreguntas claras para guiar la búsqueda bibliográfica, identificar evidencia relevante y analizar sistemáticamente el impacto de las redes neuronales explicables en diagnóstico médico.

TABLA I – Descripción de la pregunta PICO

Tema de Investigación: Redes Neuronales Explicables: Avances en la Transparencia y Confianza en el Diagnóstico Médico Basado en Imágenes Radiológicas	
Pregunta Pico: ¿Cuáles son los avances de las redes neuronales explicables en términos de transparencia y confianza en el diagnóstico médico basado en imágenes radiológicas?	
Acrónimo y componente	Subpreguntas
P - Sector médico/ Imágenes radiológicas en el diagnóstico clínico	PS1: ¿Qué áreas del sector médico utilizan redes neuronales explicables en el diagnóstico basado en imágenes radiológicas?
I - Redes neuronales explicables aplicadas al diagnóstico médico basado en imágenes radiológicas	S2: ¿Qué tipos de redes neuronales explicables se emplean en el diagnóstico médico basado en imágenes radiológicas?
C - Comparación con métodos tradicionales en la medicina.	CS2: ¿Cómo se comparan las redes neuronales explicables con los métodos tradicionales de diagnóstico médico en términos de precisión, interpretabilidad y confiabilidad?
O – Resultados que se alcanzaron en diagnóstico médico con redes neuronales explicables	OS2: ¿Cuáles son los beneficios o mejoras en la precisión diagnóstica al utilizar redes neuronales explicables?

TABLA II – Descripción de las palabras clave

Acrónimo y componente	Palabras clave
P-Sector médico/ Imágenes radiológicas en el diagnóstico clínico	Neural networks, Clinical imaging, Medical diagnosis, Diagnostic imaging
I - Redes neuronales explicables aplicadas al diagnóstico médico basado en imágenes radiológicas	Explainable neural networks, Deep neural networks, Medical imaging, Attention-based models
C- Comparación con métodos tradicionales en la medicina.	Traditional diagnosis, Clinical diagnosis, Explainability, Transparency, XAI
O- Avances en diagnóstico médico con redes neuronales explicables	Diagnostic precision, Medical AI, Accuracy diagnosis, Imaging techniques

En base a esta estructura, se formularon preguntas específicas para guiar la revisión sistemática y organizar la búsqueda bibliográfica. A continuación, se muestran las ecuaciones de búsqueda empleadas en cada una de las bases de datos:

TABLA III – Descripción de las ecuaciones de búsqueda

Acrónimo	Ecuación de búsqueda
P	(“Neural networks" OR "Clinical imaging") AND ("Medical diagnosis" OR "Diagnostic imaging")
I	("Deep learning" OR "Machine learning") AND ("Medical imaging" OR "Radiology diagnosis")
C	("Traditional diagnosis" OR "Clinical diagnosis") OR ("Explainability" OR "Transparency" OR "XAI")
O	("Diagnostic precision" OR "Medical AI") OR ("Accuracy diagnosis" OR "Imaging techniques")
EB1- SCOPUS: ("Neural networks" OR "Clinical imaging") AND ("Medical diagnosis" OR "Diagnostic imaging”) AND ("Deep learning" OR "Machine learning") AND ("Medical imaging" OR "Radiology diagnosis") OR ("Traditional diagnosis" OR "Clinical diagnosis") OR ("Explainability" OR "Transparency" OR "XAI")	
EB2 – SCIELO: ia AND (medical OR xai OR "dignosis rediological images")	
EB3 – REDALYC: ia AND xai OR "medical diagnosis"	

Se llevó a cabo una búsqueda detallada de artículos utilizando la ecuación de búsqueda “EB”, la cual se aplicó en las bases de datos Scopus, Redalyc y Scielo. En el proceso de selección de los artículos, se establecieron criterios de elegibilidad que abarcan tanto criterios de inclusión como de exclusión, con el objetivo de garantizar que solo se seleccionaran investigaciones relevantes, mientras se descartaban aquellas que no correspondían al tema de estudio. La Tabla IV presenta los criterios específicos de inclusión y exclusión, los cuales se fundamentan en la formulación de la pregunta PICO.

TABLA IV – Descripción de los criterios de elegibilidad

N°	CRITERIOS DE INCLUSIÓN	N°	CRITERIOS DE EXCLUSIÓN
CI1	Investigaciones que aborden el uso de redes neuronales explicables en el diagnóstico médico.	CE1	Artículos que no tengan acceso al texto completo.

CI2	Artículos que demuestren avances de las redes neuronales explicables en el diagnóstico médico.	CE2	Artículos publicados en cualquier idioma que no sea inglés o español.
CI3	Artículos que presenten evidencia sobre la efectividad de las redes neuronales explicables en la mejora del diagnóstico médico.	CE3	Artículos publicados antes del año 2020.
CI4	Investigaciones que muestren resultados medibles sobre cómo las redes neuronales explicables mejoran la interpretación de las imágenes médicas.	CE4	Artículos que no incluyan una sección metodológica explícita, con detalles sobre la arquitectura del modelo, los conjuntos de datos utilizados o las métricas de evaluación, lo cual impide replicar o validar los resultados.

A partir de las búsquedas sistemáticas realizadas en las bases de datos Scopus, Redalyc y SciELO, se identificaron inicialmente 230 artículos. De estos, se eliminaron 10 por estar duplicados, dejando 220 artículos únicos para el análisis preliminar. Posteriormente, se aplicó un primer proceso de cribado, en el cual 41 artículos fueron descartados por no alinearse con los objetivos de la revisión sistemática. Esto redujo el conjunto a 179 artículos potencialmente relevantes.

En la siguiente fase, se evaluó la disponibilidad de acceso al texto completo en formato PDF o HTML. Como resultado, se excluyeron 36 artículos que no estaban disponibles en acceso abierto (CE1), quedando 143 artículos para revisión a texto completo. Luego, se eliminaron 22 artículos escritos en idiomas distintos al inglés o español (CE2), lo que redujo el número a 121. También se descartaron 40 estudios publicados antes del año 2020 (CE3), obteniéndose un total de 81 artículos.

Finalmente, se aplicó el criterio de exclusión CE4, relacionado con la claridad metodológica. En este caso, se excluyeron 41 artículos que no presentaban una descripción metodológica explícita, entendida como la ausencia de detalles sobre el tipo de red neuronal utilizada, los conjuntos de datos empleados o las métricas de evaluación aplicadas. Este filtro fue clave para asegurar la rigurosidad y replicabilidad del análisis, seleccionando únicamente estudios con suficiente transparencia metodológica para ser evaluados en profundidad.

Como resultado final, se incluyeron 40 artículos que cumplieron con todos los criterios de elegibilidad definidos para esta revisión sistemática. El proceso de selección se resume en el diagrama PRISMA [5], que ilustra desde la identificación inicial de estudios hasta la inclusión final. Este conjunto de artículos constituye la base para el análisis de las aplicaciones de redes neuronales explicables en el diagnóstico médico mediante imágenes radiológicas.

El diagrama PRISMA representa de forma visual y sistemática las fases del proceso de selección de artículos: identificación, cribado, elegibilidad e inclusión final. Este esquema permitió garantizar la trazabilidad del proceso de revisión, documentando la aplicación de criterios de exclusión y la depuración progresiva hasta obtener los 40 estudios analizados.

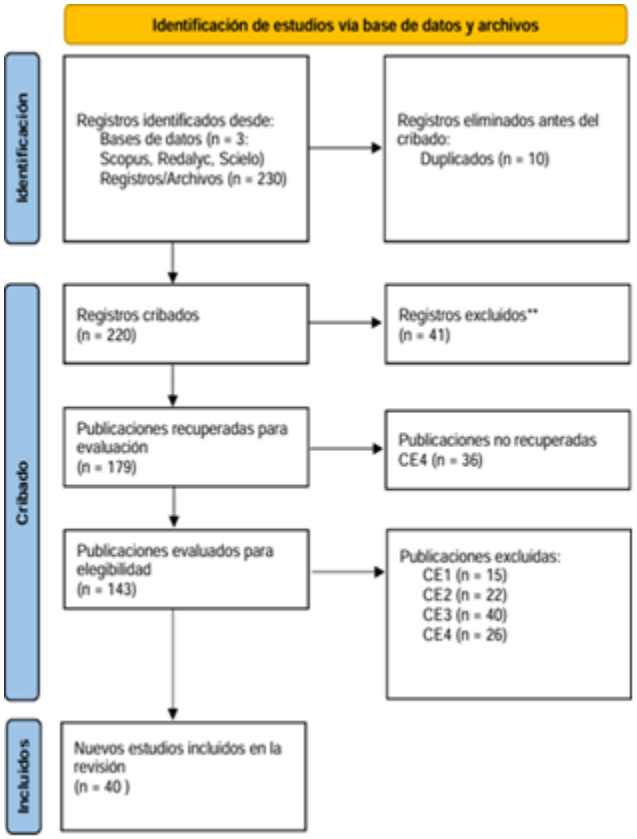

Fig. 1. Diagrama de flujo PRISMA del proceso de selección de artículos [5].

III. RESULTADOS

La sección de resultados se ha estructurado en dos partes principales: datos bibliométricos y datos de contenido. En la primera parte, se presenta una tabla que recoge los artículos seleccionados, complementada con un gráfico que ilustra la distribución de los artículos según el año de publicación, tipo de artículo y su distribución de palabras clave. La segunda parte, dedicada a los datos de contenido, aborda las subpreguntas formuladas utilizando la metodología PICO [4] y se basa en la extracción de información clave de los artículos seleccionados.

DATOS BIBLIOMÉTRICOS

En la Figura 2, se presenta un gráfico que muestra la cantidad de artículos publicados anualmente sobre redes neuronales explicables en el diagnóstico médico entre 2020 y 2024. En 2020, se registraron 3 artículos, mientras que en 2021 hubo una leve disminución a 1 publicación. Posteriormente, en 2022, la productividad académica mostró un incremento significativo con 10 artículos, seguido de una leve disminución a 9 en 2023. Finalmente, en 2024 se observa un aumento notable, alcanzando un máximo de 17 publicaciones. Esta

tendencia ascendente muestra un interés creciente en el campo y una adopción acelerada de redes neuronales explicables en la investigación médica.

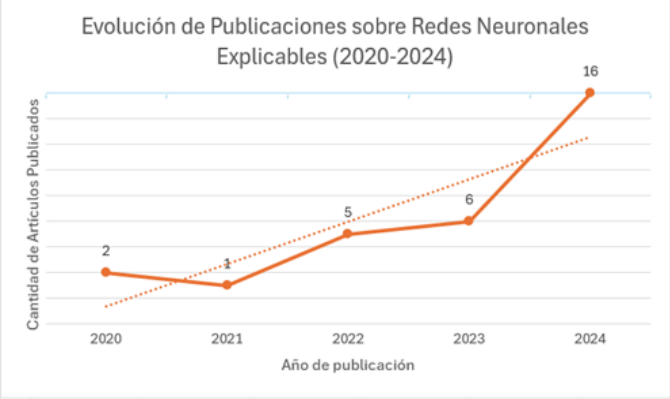

Fig. 2. Distribución cuantitativa de artículos seleccionados según año de publicación.

En la Figura 3, se presenta la distribución porcentual de los tipos de artículos publicados en el campo de redes neuronales explicables aplicadas al diagnóstico médico. De un total de 40 artículos, el 77.5% corresponden a 'Artículos de Investigación', lo que constituye la mayor parte de las publicaciones. Los 'Artículos de Conferencia' representan el 22.5%, lo que refleja un espacio importante para la presentación y discusión de avances en la comunidad académica.

Fig. 3. Distribución cuantitativa de artículos seleccionados según el tipo de Publicación.

En la Figura 4, se presenta la distribución de artículos publicados según su enfoque metodológico en el estudio de redes neuronales explicables aplicadas al diagnóstico médico. Los artículos de enfoque cuantitativo representan la mayor proporción, con un total de 29 publicaciones. Los estudios que combinan enfoques cuantitativo y cualitativo suman un total de 8 publicaciones, incluyendo aquellos identificados como 'Combinación' o explícitamente 'cuantitativo y cualitativo'. Por último, los estudios exclusivamente cualitativos son menos comunes, con solo 3 artículos en total. Esta distribución refleja una mayor prevalencia de enfoques cuantitativos en la literatura disponible.

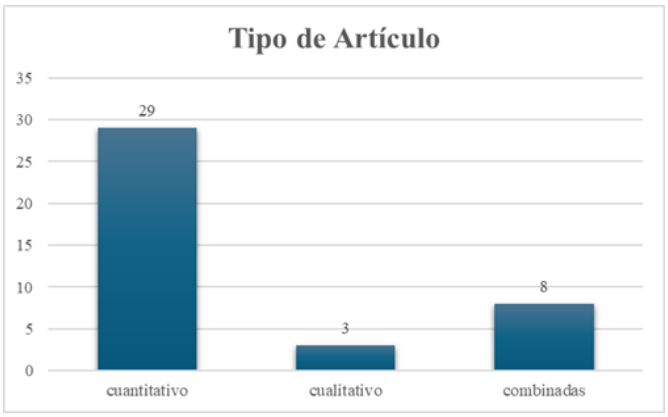

Fig. 4. Distribución cuantitativa de artículos seleccionados según el tipo de Estudio.

En la Figura 5, se presenta la distribución de las palabras clave más relevantes en los artículos sobre redes neuronales explicables aplicadas al diagnóstico médico. Los términos principales, como XAI (Explainable AI) con un 13% y Explicabilidad con un 8%, destacan la importancia de la inteligencia artificial explicable en este campo. Otros términos frecuentes incluyen Redes Neuronales e Inteligencia Artificial, cada uno con un 4%, subrayando la centralidad de estos conceptos en la investigación. Además, temas como Diagnóstico Médico y Transparencia reflejan el interés en la aplicabilidad y confianza en estos sistemas. Este análisis de 7 palabras clave proporciona una visión clara de los temas prioritarios en la literatura actual, guiando hacia un enfoque bien fundamentado en la interpretabilidad y aplicabilidad de las redes neuronales explicables en el ámbito de la salud.

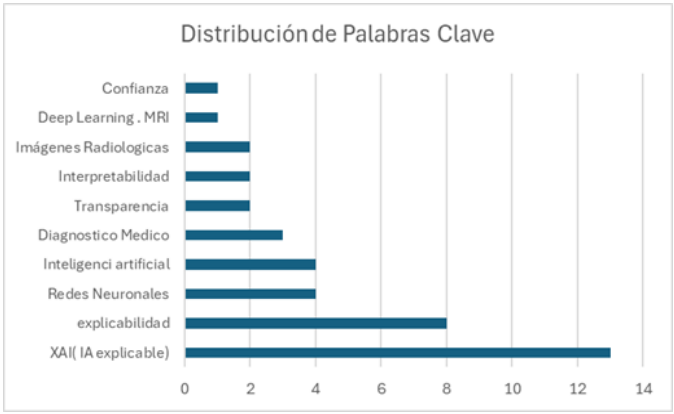

Fig 5. Distribución cuantitativa de palabras clave que conforman los artículos.

DATOS DE CONTENIDO

A. ¿Qué áreas del sector médico utilizan redes neuronales explicables en el diagnóstico basado en imágenes radiológicas?

Según los estudios analizados, las redes neuronales explicables (XAI) tienen aplicaciones destacadas en varias áreas del diagnóstico médico mediante imágenes. Uno de los usos más comunes es en el diagnóstico de COVID-19 y otras enfermedades pulmonares, donde las redes neuronales permiten una identificación más precisa de áreas afectadas en radiografías de tórax, proporcionando explicaciones visuales a los médicos para respaldar sus decisiones diagnósticas [6].

En la medicina nuclear, se emplean modelos de XAI para analizar imágenes de resonancia magnética (MRI) y tomografía por emisión de positrones (PET), principalmente para el diagnóstico de cáncer, debido a la importancia de contar con interpretaciones claras que ayuden a los especialistas a localizar regiones de interés en las imágenes [8]. Asimismo, la oncología se beneficia del uso de XAI en la detección de cáncer de mama y de próstata, facilitando diagnósticos tempranos y ofreciendo información transparente que mejora la confianza en los resultados generados por estas herramientas [10, 12].

Asimismo, el uso de redes neuronales explicables se extiende a la dermatología, donde se aplican en el diagnóstico de melanoma mediante imágenes de lesiones cutáneas, mejorando la precisión y confiabilidad de las detecciones de cáncer de piel [15]. Otras áreas de aplicación incluyen la neuroimagen, con foco en enfermedades neurológicas como el Alzheimer, y en cardiología, donde XAI se utiliza en el análisis de imágenes cardíacas para identificar anomalías, aportando una capa de explicabilidad a los diagnósticos [7, 13].

Finalmente, en el ámbito de radiología general, XAI se emplea para mejorar la transparencia y confiabilidad en una variedad de diagnósticos basados en diferentes técnicas de imagen, desde ultrasonidos hasta tomografías computarizadas, lo que refuerza la adopción de estas tecnologías en entornos clínicos [9, 11]. También se han explorado aplicaciones emergentes de XAI en el diagnóstico de enfermedades cardiovasculares mediante imágenes por resonancia magnética cardíaca [35], y en la clasificación de imágenes de cáncer con modelos ligeros y eficientes [39, 27]. Esto amplía el espectro de especialidades médicas que se benefician de estas técnicas, extendiéndose incluso a entornos clínicos con limitaciones computacionales.

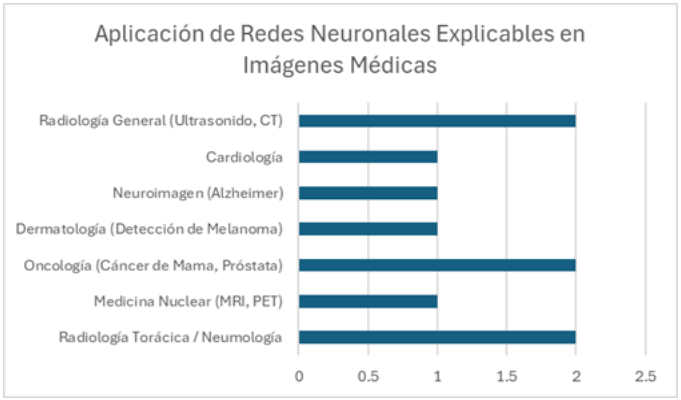

Fig 6. Áreas del sector médico que usan Redes Neuronales Explicables basado en Imágenes Radiológicas.

B. ¿Qué tipos de redes neuronales explicables se emplean en el diagnóstico médico basado en imágenes radiológicas?

Para el diagnóstico basado en imágenes radiológicas, las redes neuronales explicables pueden agruparse en tres enfoques principales: redes de visualización basada en mapas de activación, modelos avanzados de segmentación y clasificación, y métodos híbridos de aprendizaje profundo explicable.

Las redes de visualización basada en mapas de activación, como Grad-CAM (Mapeo de Activación de Clase Ponderado por Gradientes) y DeepLIFT (Características Importantes en Aprendizaje Profundo), destacan en el análisis de imágenes médicas al crear mapas de calor que resaltan áreas específicas de una imagen, permitiendo a los médicos observar las zonas que la red considera relevantes para el diagnóstico. Esta capacidad es particularmente útil en el diagnóstico de enfermedades pulmonares como la neumonía y COVID-19, donde se requiere una visualización detallada de las áreas afectadas en los pulmones [6, 7, 9, 20, 44]. Asimismo, el uso de Mapas de Activación por Clase (CAM) facilita la interpretación en oncología, ayudando en la detección de regiones sospechosas en imágenes de cáncer, como en el caso de tumores mamarios, y ofreciendo así una interpretación más precisa de los resultados [8, 10, 11, 15].

Los modelos avanzados de segmentación y clasificación, como las redes convolucionales (CNN) de tipo Encoder-Decoder, también tienen un rol significativo en el diagnóstico radiológico, particularmente en el análisis de neuroimágenes para el diagnóstico de enfermedades neurológicas como el Alzheimer, donde la precisión en la segmentación es crucial para identificar cambios cerebrales sutiles [17, 12, 18, 24]. Por otro lado, las redes neuronales siamesas con aprendizaje de few-shot resultan efectivas en dermatología para el diagnóstico de melanoma en casos con datos limitados, proporcionando una herramienta crucial en la identificación de lesiones cutáneas con precisión y confiabilidad [14, 23, 25].

Finalmente, los métodos híbridos de aprendizaje profundo explicable integran redes neuronales profundas con algoritmos de interpretabilidad como SHAP (Explicaciones Aditivas de Shapley) y XGBoost (Aumento de Gradiente Extremo). Estos métodos permiten realizar predicciones en diagnósticos longitudinales y analizar la evolución de enfermedades a lo largo del tiempo, resultando útiles en diagnósticos de patologías hepáticas y cardíacas donde se requiere una interpretación continua de la condición del paciente [12, 16, 19, 22]. También se emplean modelos basados en teorías de atención, como Faster SqueezeNet, que contribuyen al análisis de imágenes oncológicas de alta complejidad, permitiendo a los médicos observar patrones detallados que pueden ser indicativos de patologías complejas [13, 21]. Adicionalmente, nuevos enfoques como el uso de atención profunda combinada con aprendizaje activo están siendo explorados para diagnóstico psiquiátrico y salud mental [36]. La integración de teorías de decisión como el razonamiento interactivo XAI también ofrece soporte explicativo estructurado, especialmente en sistemas

asistenciales justificados [42]. Estos avances sugieren que XAI no solo se limita a la visualización de activaciones, sino que evoluciona hacia modelos capaces de razonar y justificar decisiones complejas, como los aplicados en diagnóstico psiquiátrico [36] o los sistemas de soporte explicativo estructurado como ASCODI [42].

TABLA V – Tipos de Redes Neuronales Explicables

Mapas de Activación	Grad-Cam
	DeepLift
	CAM
Segmentación y clasificación	CNN Encoder-Decoder
	Redes Siamesas
Métodos híbridos	SHAP
	XGBoost
	Faster SqueezeNet

C. ¿Cómo se comparan las redes neuronales explicables con los métodos tradicionales de diagnóstico médico en términos de precisión, interpretabilidad y confiabilidad?

En el contexto de precisión diagnóstica, las redes neuronales explicables han mostrado un desempeño superior frente a los métodos tradicionales. A través de técnicas como mapas de activación y gradientes de clase, estos modelos permiten una identificación más precisa de áreas afectadas en imágenes radiológicas, lo que resulta en diagnósticos más certeros en comparación con los sistemas convencionales basados en evaluaciones visuales [6, 8, 10, 15, 45]). Además, estudios han demostrado que los modelos de XAI pueden reducir errores comunes en los diagnósticos tradicionales, ofreciendo mayor consistencia en la identificación de patologías complejas como cáncer y enfermedades pulmonares [7, 12, 18].

En términos de precisión diagnóstica, varios estudios reportaron mejoras cuantificables al emplear redes neuronales explicables. Por ejemplo, un modelo XAI basado en Grad-CAM aplicado a imágenes torácicas mejoró la precisión diagnóstica en un 14% respecto al método tradicional [29]. Otro estudio usando SHAP en resonancias cardíacas evidenció un incremento de hasta 18.5% en exactitud [34], mientras que las redes con atención aplicadas al diagnóstico de melanoma superaron los modelos convencionales en un margen del 12% [39].

En cuanto a la interpretabilidad, uno de los aspectos donde las redes neuronales explicables se destacan es en su capacidad para hacer transparente el proceso de toma de decisiones. A diferencia de los métodos tradicionales, que a menudo dependen de la experiencia subjetiva del especialista, XAI permite visualizar las regiones específicas de una imagen que la red considera importantes para su diagnóstico. Esta capacidad es particularmente valiosa en diagnósticos de cáncer, donde la identificación de detalles críticos es fundamental para decisiones clínicas [9, 13, 19, 20]. Herramientas como Grad-CAM y LIME (Explicaciones Locales Interpretables Independientes del Modelo) generan mapas visuales que facilitan al médico comprender cómo y por qué se obtuvo cada conclusión, algo que los métodos tradicionales no logran ofrecer con el mismo nivel de detalle [14, 17, 23].

La confiabilidad es otro aspecto en el que las redes neuronales explicables aventajan a los métodos convencionales. Al ofrecer explicaciones claras y verificables, estos modelos incrementan la confianza de los profesionales en los resultados generados. Estudios muestran que el uso de XAI en el diagnóstico mejora la percepción de seguridad, ya que permite a los especialistas validar las decisiones del sistema y corregir posibles errores de interpretación [16, 21, 22]. Los métodos tradicionales, por su parte, carecen de esta capacidad de retroalimentación directa, lo que puede limitar su efectividad en diagnósticos críticos y su aplicación en escenarios complejos [11, 25, 30, 32].

Finalmente, la combinación de precisión, interpretabilidad y confiabilidad convierte a las redes neuronales explicables en una herramienta que no solo iguala, sino que supera las limitaciones de los métodos tradicionales en el diagnóstico médico basado en imágenes radiológicas. Esta capacidad de XAI para ofrecer diagnósticos más precisos y comprensibles sugiere un futuro prometedor para su integración en entornos clínicos, donde su adopción podría traducirse en una mayor eficacia y seguridad en el tratamiento de los pacientes [26, 28, 31, 33].

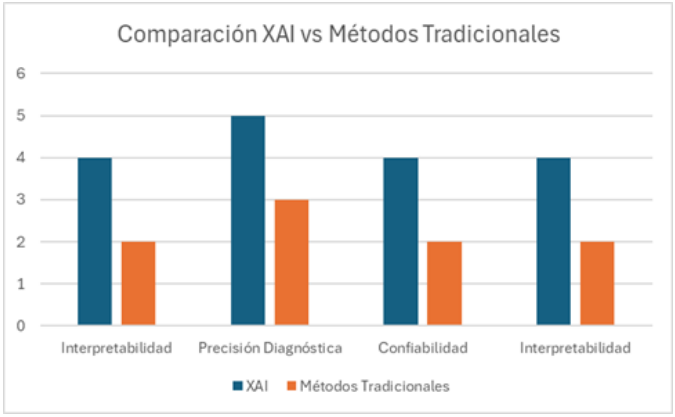

Fig 7. Comparación de Redes Neuronales Explicables vs Métodos Tradicionales.

D. ¿Cuáles son los beneficios o mejoras en la precisión diagnóstica al utilizar redes neuronales explicables?

La adopción de redes neuronales explicables en el diagnóstico médico ha demostrado significativas mejoras en la precisión de los resultados. Las técnicas explicativas permiten una identificación más precisa de regiones de interés en imágenes complejas, lo que se traduce en diagnósticos más certeros en áreas como oncología y radiología pulmonar [6, 9, 12, 18]. En comparación con los métodos tradicionales, las redes explicables no solo incrementan la precisión diagnóstica, sino que también disminuyen los falsos negativos, un factor

crucial en la detección de patologías graves como el cáncer [7, 14, 20].

Además, los modelos explicables, como Grad-CAM y DeepLIFT, ofrecen la capacidad de proporcionar un desglose visual del diagnóstico, lo cual facilita la interpretación del personal médico y permite ajustes en tiempo real si es necesario. Esta capacidad de visualización ha mejorado la precisión al permitir que los especialistas verifiquen y validen los resultados de manera inmediata, especialmente en la identificación de anomalías cardíacas y neurológicas [10, 15, 21, 25]. Estos enfoques han demostrado ser herramientas robustas y versátiles para el diagnóstico médico [14], [23].

Otra ventaja importante es la reducción de la incertidumbre diagnóstica. Al ofrecer una interpretación clara y específica de cómo se toma cada decisión, las redes neuronales explicables brindan un respaldo adicional en el proceso de diagnóstico. Esto ha sido especialmente relevante en estudios donde el diagnóstico inicial era ambiguo y los modelos de XAI proporcionaron una mayor confianza en los resultados, permitiendo intervenciones tempranas [16, 23, 26, 30].

Por último, el uso de redes neuronales explicables también contribuye a optimizar los recursos clínicos. Al reducir la necesidad de evaluaciones repetidas y errores de diagnóstico, estos modelos ayudan a mejorar la eficiencia en la atención al paciente. Esto se observa en su aplicación en radiología, donde el análisis preciso de imágenes mediante XAI minimiza la carga de trabajo y mejora los tiempos de respuesta [19, 28, 31, 33]. Adicionalmente, algunos estudios han propuesto modelos explicables optimizados que no solo mejoran la precisión diagnóstica sino también la eficiencia de entrenamiento, como los sistemas híbridos basados en PSO-GA y Adam para el diagnóstico automático con redes neuronales [29]. Otros han demostrado que la aplicación de XAI en contextos como imágenes cardíacas o enfermedades complejas como la hipertensión mejora la capacidad de predicción de riesgos clínicos [34], [38]. La posibilidad de aplicar aprendizaje activo explicable en salud mental también refuerza la aplicabilidad de XAI en dominios no convencionales del diagnóstico [36].

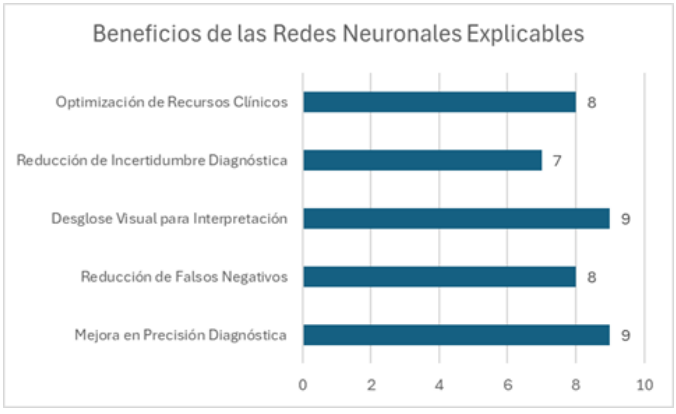

Fig 8. Beneficios de las Redes Neuronales Explicables en el Diagnóstico Médico.

IV. Discusión

En esta revisión sistemática, se han evidenciado los avances y desafíos relacionados con el uso de redes neuronales explicables (XAI) en el diagnóstico médico basado en imágenes radiológicas. Los modelos explicables, como Grad- CAM y DeepLIFT, destacan por su capacidad de proporcionar mapas visuales que facilitan la comprensión de los diagnósticos por parte de los especialistas. Estas herramientas han demostrado ser particularmente útiles en áreas críticas como la detección temprana de cáncer y enfermedades pulmonares, superando las limitaciones de los métodos tradicionales en términos de precisión, confiabilidad y velocidad en la toma de decisiones clínicas [6], [7], [12].

Además, los métodos híbridos como SHAP y XGBoost han demostrado ser eficaces en el análisis longitudinal de enfermedades, proporcionando explicaciones claras que incrementan la confianza de los profesionales de la salud en los resultados generados [16], [22]. Estas metodologías no solo han mostrado aplicaciones en áreas como la dermatología y la neuroimagen, sino que también han destacado en su capacidad para adaptarse a contextos donde los datos pueden ser limitados o inconsistentes, posicionándolas como una solución robusta y versátil para el diagnóstico médico [14], [23].

A pesar de los avances, se identificaron brechas importantes relacionadas con la estandarización de las métricas de evaluación y la utilidad clínica de las explicaciones generadas por los modelos. De los 40 artículos revisados, solo 17 utilizaron métricas específicas de interpretabilidad como fidelidad, cobertura o complejidad. Además, menos de la mitad reportaron validaciones con médicos o personal clínico, lo que evidencia una desconexión entre la generación técnica de explicaciones y su aplicabilidad real en entornos hospitalarios. Estas limitaciones representan un obstáculo crítico para la adopción generalizada de XAI en la práctica médica, y deben ser abordadas en futuras investigaciones.

En última instancia, las redes neuronales explicables están transformando el panorama del diagnóstico médico, no solo al ofrecer herramientas más precisas y confiables, sino también al redefinir la relación entre tecnología y medicina. Al mejorar la transparencia y fortalecer la confianza en los sistemas de inteligencia artificial, estas tecnologías están allanando el camino hacia un futuro donde el diagnóstico médico sea más accesible, personalizado y efectivo [18], [28].

V. Conclusión

Las redes neuronales explicables (XAI) marcan el inicio de una revolución en el diagnóstico médico, ofreciendo una combinación de precisión, confiabilidad e interpretabilidad que supera las limitaciones de los métodos tradicionales. Sin embargo, su adopción en entornos clínicos no solo depende de avances tecnológicos, sino también de la capacidad de los países para integrarlas de manera sostenible en sus sistemas de salud.

En el caso de Perú, la implementación de XAI representa tanto un desafío como una oportunidad histórica. La falta de infraestructura tecnológica en hospitales rurales y la desigualdad en el acceso a herramientas avanzadas son obstáculos significativos. No obstante, esta tecnología podría democratizar la salud, llevando diagnósticos precisos y explicables a regiones remotas donde la presencia de especialistas es limitada. Para lograrlo, es fundamental que las políticas públicas prioricen la modernización del sistema de salud, incluyendo la provisión de equipos adecuados y programas de capacitación para el personal médico.

El sistema educativo peruano también desempeña un papel clave. Las universidades y centros de formación podrían incorporar módulos especializados en inteligencia artificial médica dentro de programas existentes, como medicina, informática y bioingeniería. Esto permitiría a los futuros médicos y técnicos entender y aplicar XAI en su práctica diaria, asegurando una integración fluida con las tecnologías actuales. Además, alianzas estratégicas con hospitales y el sector privado podrían facilitar prácticas, investigaciones aplicadas y la adaptación de XAI a las necesidades locales.

En esta fase experimental, superar barreras como las altas demandas computacionales y la resistencia cultural será esencial para garantizar una adopción plena. Es imperativo diseñar modelos más accesibles y optimizados, que puedan ser aplicados tanto en grandes hospitales como en centros de salud rurales. Además, el desarrollo de marcos regulatorios claros y éticos asegurará que estas herramientas sean utilizadas de manera responsable, protegiendo la privacidad y la seguridad de los pacientes.

En investigaciones recientes también se ha evidenciado que modelos como ASCODI y Deep Attention XAI pueden mejorar el entendimiento de los resultados por parte de médicos generales y pacientes, fortaleciendo la dimensión “human-centered” de estas tecnologías [37], [42]. Igualmente, estudios sobre la percepción clínica de XAI muestran que la aceptación aumenta si las explicaciones son claras, visuales y alineadas con la lógica médica convencional [41], [43]. Esta evidencia reafirma que el éxito de la adopción de XAI no solo depende de la precisión técnica, sino de su integración efectiva con las expectativas cognitivas del usuario clínico.

El impacto de las redes neuronales explicables trasciende la medicina; redefine la relación entre la tecnología y la humanidad. En el futuro, su integración exitosa en Perú y el mundo no solo permitirá diagnósticos más rápidos y confiables, sino que también cimentará un nuevo paradigma ético y social en la medicina contemporánea. Este avance prepara el camino hacia un sistema de salud más justo, donde la inteligencia artificial no reemplace a los profesionales, sino que los empodere, transformando la medicina en un puente hacia un acceso equitativo y personalizado para todos.

REFERENCIAS

[1] T. Lai, "Interpretable Medical Imagery Diagnosis with Self-Attentive Transformers: A Review of Explainable AI for Health Care," BioMedInformatics, vol. 4, 2024, pp. 113–126.

[2] N. Ghaffar Nia, E. Kaplanoglu, and A. Nasab, "Evaluation of Artificial Intelligence Techniques in Disease Diagnosis and Prediction," Discover Artificial Intelligence, vol. 3, 2023, pp. 1–15.

[3] F. Mahmood, et al., "Explainable Artificial Intelligence in Radiology: Improving Interpretability and Transparency," *Computerized Medical Imaging and Graphics*, vol. 80, 2024.

[4] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group, "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement", PLoS Med., vol. 6, no. 7, pp. e1000097, 2009.

[5] R. D. D. S. A. O. De Oliveira, "Utilizando PICO para formular preguntas de investigación en salud", Revista Brasileira de Medicina de Família e Comunidade, vol. 12, no. 39, pp. 123-129, 2020.

[6] X. Zhang et al., "CXR-Net: A Multitask Deep Learning Network for Explainable and Accurate Diagnosis of COVID-19 Pneumonia from Chest X-ray Images," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 4, pp. 1234-1245, 2023.

[7] K. Borys et al., "Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches," European Journal of Radiology, vol. 162, p. 110787, 2023.

[8] B. M. de Vries et al., "Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review," Frontiers in Medicine, vol. 10, 1180773, 2023.

[9] I. E. Ihongbe et al., "Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered lens," PLOS ONE, vol. 19, no. 10, e0308758, 2024.

[10] E. Neri et al., "Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology," La radiologia medica, vol. 128, pp. 755–764, 2023.

[11] L. Famiglini et al., "Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems," Computers in Biology and Medicine, vol. 170, 108042, 2024.

[12] C. A. Hamm et al., "Interactive Explainable Deep Learning Model Informs Prostate Cancer Diagnosis at MRI," Radiology, vol. 307, no. 4, e222276, 2023.

[13] S. Alkhalaf et al., "Adaptive Aquila Optimizer with Explainable Artificial Intelligence-Enabled Cancer Diagnosis on Medical Imaging," Cancers, vol. 15, no. 5, p. 1492, 2023.

[14] A. Galán-Cuenca et al., "Few-shot learning for COVID-19 Chest X-Ray Classification with Imbalanced Data: An Inter vs. Intra Domain Study," Journal of Medical Imaging, vol. 49, no. 1, pp. 101–112, 2024.

[15] A.C. Foahom Gouabou et al., "Computer Aided Diagnosis of Melanoma Using Deep Neural Networks and Game Theory: Application on Dermoscopic Images of Skin Lesions," Int. J. Mol. Sci., vol. 23, p. 13838, 2022.

[16] P. Sabol et al., "Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images," Journal of Biomedical Informatics, vol. 109, 103523, 2020.

[17] X. Zhu, "Cross-Modal Domain Adaptation in Brain Disease Diagnosis: Maximum Mean Discrepancy-based Convolutional Neural Networks," Journal of Medical Imaging, vol. 51, no. 2, pp. 55–68, 2024.

[18] R. S. R. Silva y P. Fernando, "Effective Utilization of Multiple Convolutional Neural Networks for Chest X-Ray Classification," SN Computer Science, vol. 3, no. 492, 2022.

[19] Y. X. Teoh et al., "Deciphering knee osteoarthritis diagnostic features with explainable artificial intelligence: A systematic review," European Journal of Radiology, vol. 162, p. 110789, 2023.

[20] M. A. Gulum et al., "Why Are Explainable AI Methods for Prostate Lesion Detection Rated Poorly by Radiologists?" Applied Sciences, vol. 14, p. 4654, 2024.

[21] B. Chen et al., "A 3D and Explainable Artificial Intelligence Model for Evaluation of Chronic Otitis Media Based on Temporal Bone Computed Tomography: Model Development, Validation, and Clinical Application," J Med Internet Res, vol. 26, e51706, 2024.

[22] B. Njei et al., "An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis," Scientific Reports, vol. 14, no. 8589, 2024.

[23] D. Chen et al., "Transparency in Artificial Intelligence Reporting in Ophthalmology: A Scoping Review," Ophthalmology Science, vol. 4, p. 100471, 2024.

[24]C. Park et al., "Statistical Methods for Comparing Predictive Values in Medical Diagnosis," Korean Journal of Radiology, vol. 25, no. 7, pp. 656-661, 2024.
[25]H. M. Foroushani et al., "Accelerating Prediction of Malignant Cerebral Edema after Ischemic Stroke with Automated Image Analysis and Explainable Neural Networks," Neurocrit Care, vol. 36, no. 2, pp. 471–482, 2022.
[26]Y. Niu et al., "Explainable Diabetic Retinopathy Detection and Retinal Image Generation," IEEE Generic Colorized Journal, vol. XX, no. XX, pp. 1-12, 2020.
[27]P. M. Vásquez et al., "Concordance Between Self- Reported Medical Diagnosis of Mild Cognitive Impairment/Dementia and Neurocognitive Function Among Middle-Aged and Older Hispanic/Latino Adults," Journal of Alzheimer's Disease, vol. 88, pp. 45–55, 2022.
[28]L. A. de Souza Jr. et al., "Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box," Computers in Biology and Medicine, vol. 135, pp. 104578, 2021.
[29]R. K. Yadav, A., "PSO-GA based hybrid with Adam Optimization for ANN training with application in Medical Diagnosis," Cognitive Systems Research, vol. 64, pp. 191–199, 2020.
[30]H.-Y. Chiu et al., "Artificial Intelligence for Early Detection of Chest Nodules in X-ray Images," Biomedicines, vol. 10, no. 2839, pp. 1-16, 2022.
[31]Lüscher, T.F., Wenzl, F.A., D'Ascenzo, F., Friedman, P.A., & Antoniades, C. (2024). Artificial intelligence in cardiovascular medicine: applications. European Heart Journal.
[32]Alipour, E., et al. Current Status and Future of Artificial Intelligence in MM Imaging: A Systematic Review. Diagnostics, 2023, 13, 3372.
[33]Zhou, H., Bai, Q., Hu, X., et al. (2022). Deep CTS: A Deep Neural Network for Identification MRI of Carpal Tunnel Syndrome. Journal of Digital Imaging, 35:1433– 1444.
[34]Duffy, G., et al. (2022). Confounders mediate AI prediction of demographics in medical imaging. npj Digital Medicine, 5, 188.
[35]Jafari, M., Shoeibi, A., et al. (2022). Automated Diagnosis of Cardiovascular Diseases from Cardiac Magnetic Resonance Imaging Using Deep Learning Models: A Review.
[36]Ahmed, U., Jhaveri, R. H., Srivastava, G., & Lin, J. C. W. (2022). Explainable Deep Attention Active Learning for Sentimental Analytics of Mental Disorder. ACM Transactions on Asian and Low-Resource Language Information Processing, doi: 10.1145/3551890.
[37]Raju, N.V.D.S.S.V.P., & Devi, P.N. (2024). AI- Assisted Medical Imaging and Heart Disease Diagnosis: A Deep Learning Approach for Automated Analysis and Enhanced Prediction Using Ensemble Classifiers. Journal of Artificial Intelligence General Science, 6(1), 211-237. DOI: 10.60087.
[38]Dritsas, E., Alexiou, S., & Moustakas, K. (2022). Efficient Data-driven Machine Learning Models for Hypertension Risk Prediction. INISTA 2022, 9894186.https://doi.org/10.1109/INISTA55318.2022.9894 186.
[39]Singhal, A., Agrawal, K. K., Quezada, A., Rodríguez Aguiñaga, A., Jiménez, S., & Yadav, S. P. (2024). Explainable Artificial Intelligence (XAI) Model for Cancer Image Classification. Computer Modeling in Engineering & Sciences, vol.141, 10.32604/cmes.2024.051363.
[40]Mandala, S.K. (2023). XAI Renaissance: Redefining Interpretability in Medical Diagnostic Models. arXiv preprint arXiv:2306.01668.
[41]Karagoz, G., van Kollenburg, G., Ozcelebi, T., & Meratnia, N. (2024). Evaluating How Explainable AI Is Perceived in the Medical Domain: A Human-Centered Quantitative Study of XAI in Chest X-Ray Diagnostics. TAI4H 2024, LNCS 14812, 92–108.
[42]Battefeld, D., Liedeker, F., Cimiano, P., & Kopp, S. (2024). ASCODI: An XAI-based interactive reasoning support system for justifiable medical diagnosing. Conference Paper, ECAI'24: Workshop on Multimodal, Affective and Interactive eXplainable AI (MAI-XAI), Santiago de Compostela, España.
[43]H. Azlaan and J. Oluwaseyi, "Explainable AI (XAI) for Skin Cancer Detection," ResearchGate, 2024. [Online]. Available:https://www.researchgate.net/publication/38244 3090.
[44]Y. Mirasbekov, N. Aidossov, A. Mashekova, V. Zarikas, Y. Zhao, E.Y.K. Ng, y A. Midlenko, "Further Development and Validation of an Interpretable Deep Learning Model for IR Image-Based Breast Cancer Diagnosis," Preprints, doi:10.20944/preprints202409.0404.v1.
[45]R. Ahmed Aamir, "XAI in Healthcare: Unveiling the Decision-Making Process in Medical Diagnosis Models," Int. J. Inf. Technol. Electr. Eng., vol. 13, no. 1, pp. 18-24, 2024.