FACULTAD DE INFORMATICA

UNIVERSIDAD NACIONAL DE LA PLATA

“**Inteligencia Artificial Explicable**: Análisis de Metodologías y Aplicaciones”

Lic.María Cecilia Pezzini

Directora

Dra. Claudia Pons

Trabajo final integrador para obtener la especialización en ingeniería de software

La Plata, Buenos Aires, Argentina

Septiembre 2024

Índice de contenido

Capítulo 1
Presentación

"En el mundo moderno, la inteligencia artificial se ha convertido en la nueva electricidad, transformando la manera en que vivimos, trabajamos y nos relacionamos."

Andrew Ng.

1.1 Introducción

El uso de sistemas de aprendizaje automático en situaciones complejas ha generado un mayor interés en lograr su mejora, no solo para cumplir con tareas específicas, sino también para otros aspectos críticos, como garantizar la seguridad, prevenir la discriminación, evitar problemas técnicos y asegurar que los modelos puedan explicar sus decisiones de manera comprensible.

En este capítulo se presenta la introducción a la propuesta de trabajo final integrador. En la sección 1.2, se indica la motivación para el estudio de la temática planteada. La sección 1.3, detalla los objetivos de la investigación. Luego, en la sección 1.4, se expone el abordaje del trabajo de investigación. Finalmente en la sección 1.5, se describe la estructura del documento.

1.2 Motivación

En la actualidad, la inteligencia artificial (IA, por sus siglas en inglés) desempeña un papel cada vez más relevante en diversos campos (Baeza-Yates, 2024), impulsando avances significativos en la productividad, la eficiencia y la innovación. Sin embargo, este progreso también ha resaltado un desafío fundamental: la falta de transparencia en muchos sistemas de IA. En entornos críticos donde la toma de decisiones se basa en resultados generados por algoritmos de IA, la explicabilidad es esencial para garantizar la confianza, la responsabilidad y la aceptación pública.

En "The Fourth Industrial Revolution", Klaus Schwab (2017) subraya la necesidad de enfrentar los desafíos éticos que surgen con la rápida evolución tecnológica de la era actual. Destaca la necesidad de establecer marcos éticos y regulaciones apropiadas para guiar el desarrollo y uso de tecnologías emergentes. Además, Schwab enfatiza la colaboración entre empresas, gobiernos, instituciones académicas y la sociedad civil para abordar los dilemas éticos de manera efectiva.

Es en este contexto que surge la necesidad de explorar el concepto de inteligencia artificial explicable (XAI). Este trabajo se propone analizar los modelos, soluciones y desafíos actuales asociados con la IA explicable. Al hacerlo, se busca contribuir al avance de la comprensión y aplicación de la IA con un enfoque en la transparencia y la explicabilidad. Al comprender los modelos y soluciones disponibles, así como los desafíos que enfrentan, podemos promover un desarrollo responsable y ético de la IA, garantizando que se utilice de manera efectiva para el beneficio de la sociedad en su conjunto.

En línea con este propósito, surgen el siguiente interrogante:

"¿Cuáles son los avances más recientes en la mejora de la explicabilidad de los modelos de machine learning considerados 'caja negra', en comparación con enfoques anteriores, y cuál es el impacto de dichas mejoras, en términos de propuestas, taxonomías y otros resultados relevantes?"

1.3 Objetivos

1.3.1 Objetivo General

Este trabajo tiene como objetivo estudiar y analizar los principales problemas abordados en la literatura con respecto a la "Inteligencia Artificial Explicable" (XAI por sus siglas en inglés), centrándose en comprender la necesidad de transparencia y explicabilidad en los sistemas de inteligencia artificial.

Se buscará identificar los desafíos éticos y prácticos asociados con los sistemas que ocultan su lógica interna, así como explorar los diferentes enfoques desarrollados para superar esta falta de explicación.

1.3.2 Objetivos Específicos

- Revisar el concepto de inteligencia artificial explicable para comprender su importancia e implicaciones.
- Identificar el estado actual de la IA explicable mediante la sistematización del conocimiento del tema.
- Clasificar los aspectos relevantes de los sistemas de caja negra, incluyendo el problema abordado, las soluciones propuestas para la explicación, el tipo de datos analizados y el tipo de predictor explicado (Guidotti, Monreale, Ruggieri, Turini, Pedreschi y Giannotti (2018)).
- Analizar la literatura existente para identificar y categorizar los problemas fundamentales relacionados con la interpretación de sistemas de caja negra, como sesgos, falta de transparencia y falta de explicabilidad.
- Evaluar las metodologías existentes: Examinar las metodologías actuales utilizadas para abrir y comprender sistemas de caja negra, identificando sus fortalezas, debilidades y áreas de mejora.

1.4 Abordaje del trabajo

Este trabajo se basa en una revisión bibliográfica de artículos científicos y publicaciones de congresos, obtenidos de bibliotecas virtuales como ACM Digital Library, Google Scholar y IEEE Xplore. Los documentos seleccionados se presentan conforme a los criterios de búsqueda establecidos.

El proceso de revisión incluye:

- **Selección de Documentos**: Se muestran los documentos que cumplen con los criterios de inclusión definidos, asegurando su relevancia y calidad en el contexto de la Inteligencia Artificial Explicable (XAI).
- **Métodos de Análisis**: Se describen los métodos utilizados para analizar los documentos, que incluyen la evaluación de metodologías, aplicaciones prácticas y desafíos identificados.
- **Resumen y Conclusiones**: Cada documento se resume detalladamente, y se presentan las conclusiones derivadas de la revisión, proporcionando una visión crítica y consolidada de los hallazgos.

1.5 Estructura del Trabajo

El trabajo está organizado en cinco capítulos, dos anexos y una bibliografía:

Presentación Este capítulo presenta el contexto de la investigación sobre Inteligencia Artificial Explicable (XAI), planteando las preguntas de investigación y definiendo los objetivos principales. Se aborda la relevancia de la explicabilidad en los modelos de machine learning, denominados 'caja negra', y se ofrece una visión general de la estructura del trabajo.

Conceptos Teóricos Se establece una base conceptual sobre Inteligencia Artificial (IA) y XAI. Se definen conceptos clave como redes neuronales, aprendizaje supervisado y no supervisado, visión por computadora y procesamiento de lenguaje natural. También se exploran métodos de evaluación de modelos de IA, así como la importancia de la interpretabilidad y la explicabilidad.

Revisión de la Literatura Este capítulo revisa la literatura existente sobre XAI para identificar avances recientes y limitaciones. Se describe la estrategia de búsqueda bibliográfica, los criterios de inclusión y exclusión, y la metodología de análisis empleada.

Análisis de Metodologías y Aplicaciones XAI En este capítulo se exploran las metodologías y aplicaciones de la Inteligencia Artificial Explicable (XAI). Se comienza con una visión general de los estudios seleccionados (véase Tabla 2), basados en los criterios de inclusión

previamente definidos. A continuación, se realiza un análisis comparativo de diversos métodos de explicabilidad, como se presenta en las Tablas 3, 4 y 5.

El análisis se centra en los siguientes aspectos:

- **Clasificación y Comparación de Métodos**: Evaluación de las técnicas de explicabilidad según su enfoque (ante-hoc o post-hoc), su aplicabilidad a diferentes tipos de modelos (agnósticos o específicos) y su capacidad para proporcionar explicaciones globales o locales.
- **Evaluación de Aplicaciones Prácticas**: Análisis de cómo estos métodos se implementan en contextos reales, destacando casos de uso en dominios como la medicina, la visión por computadora y la toma de decisiones empresariales.
- **Impacto en la Interpretabilidad y Confianza**: Consideración de cómo cada metodología contribuye a mejorar la transparencia de los modelos y la confianza de los usuarios en las decisiones automatizadas.
- **Desafíos y Limitaciones**: Identificación de las principales limitaciones y desafíos asociados con cada enfoque, incluyendo aspectos relacionados con la escalabilidad, la complejidad computacional y la generalización de las explicaciones.
- **Tendencias Futuras**: Discusión de las tendencias emergentes y las áreas de investigación futura en XAI.

Se proporciona un marco para comprender y comparar las diversas metodologías de XAI, ofreciendo una base para la evaluación crítica de su aplicabilidad y eficacia en diferentes contextos.

Conclusiones Se resumen los hallazgos principales del estudio, reflexionando sobre sus implicaciones para la investigación y práctica futura en XAI. Se discuten las contribuciones significativas, el impacto de los avances recientes y se sugieren direcciones para futuras investigaciones. También se abordan los desafíos actuales en la mejora de la explicabilidad en modelos de IA.

Anexo I: Avances Recientes en la Explicabilidad de Modelos de IA Este anexo proporciona un análisis detallado de 30 trabajos seleccionados sobre los avances recientes en la explicabilidad de modelos de IA. Cada trabajo explora enfoques y técnicas que mejoran la comprensión y la transparencia de los modelos, especialmente en contextos críticos.

Para cada estudio se incluye:

- **Descripción de los Enfoques**: Un resumen de las metodologías y técnicas propuestas.
- **Contribuciones Clave**: Un análisis de cómo cada enfoque avanza en la explicación de modelos de IA.

- **Cuadro Resumen**: Una tabla que evalúa los aspectos relevantes de cada trabajo en relación con la pregunta de investigación.

Este análisis se utiliza para sintetizar la información en las Tablas 3, 4 y 5 del cuerpo del documento, proporcionando una visión consolidada de los avances en el campo.

Anexo II: Papers Recuperados sobre Modelos de Explicabilidad en IA Este anexo incluye todos los papers recuperados de bases de datos académicas como ACM Digital Library, Google Scholar y IEEE Xplore, utilizando palabras clave basadas en el enfoque PICO. Para cada entrada se proporciona:

- **Título del Paper**: El nombre completo del estudio.
- **Cita**: Referencia completa del paper en el formato adecuado.
- **Descripción**: Un resumen de los aspectos relevantes de cada estudio, que permite evaluar si el enfoque del paper cumple con la pregunta de investigación planteada.

Bibliografía: La sección de bibliografía incluye todas las fuentes académicas y artículos de investigación utilizados a lo largo del trabajo, presentados siguiendo el formato APA.

Capítulo 2

Conceptos Teóricos

2.1 Conceptos Fundamentales

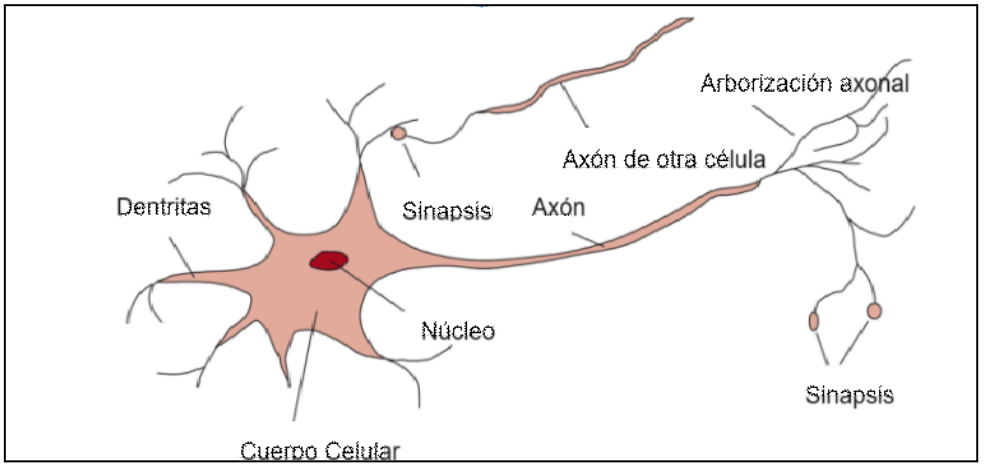
En este capítulo, se presenta el marco conceptual sobre el cual se construyó el trabajo. Se abordan los conceptos de red neuronal, aprendizaje supervisado y no supervisado, se exploran las estructuras internas de las redes neuronales convolucionales y generativas, y se examinan los desafíos y soluciones en el entrenamiento de redes adversarias. Además, se discute la evaluación de modelos, la elección de algoritmos y la cuestión de la explicabilidad e interpretabilidad de los sistemas de inteligencia artificial.

Todos los tiempos verbales están en presente, lo cual es adecuado para la descripción y exposición de conceptos en un capítulo introductorio.

2.1.1 Definición de Red Neuronal

Una red neuronal biológica es una colección de células simples que dan lugar al pensamiento, la acción y la conciencia. Como dijo John Searle (1992) de manera concisa, los cerebros causan mentes.
Cada neurona se compone de un cuerpo celular (soma), dendritas y un axón (véase Figura 1) Las sinapsis conectan las neuronas y permiten la comunicación entre ellas. Estas redes neuronales, tanto biológicas como artificiales (estas imitan el funcionamiento de las redes neuronales biológicas), son fundamentales para el procesamiento de información y el aprendizaje.
En la siguiente figura se muestran las partes de una neurona, que es la unidad fundamental del sistema nervioso (véase Figura 1).

Figura 1. Las partes de una célula nerviosa o neurona. (Russell & Norvig (2020)).

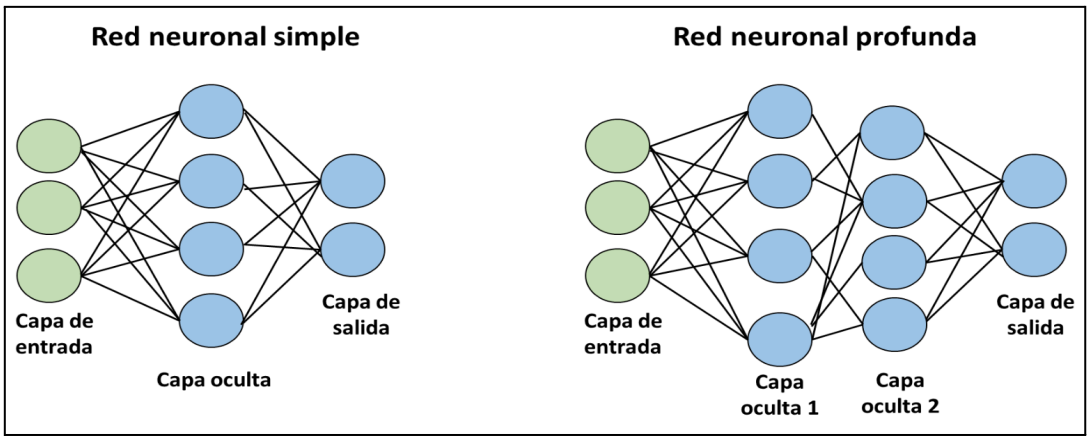


Las redes neuronales artificiales (ANN, por sus siglas en inglés, Artificial Neural Networks) imitan el funcionamiento de las redes neuronales biológicas y se utilizan en campos como la inteligencia artificial y la ciencia de la computación.

2.1.2 Redes Neuronales Profundas

Las redes neuronales feedforward, también conocidas como redes neuronales profundas o perceptrones multicapa (MLPs), son modelos fundamentales en el campo del aprendizaje profundo. Según Goodfellow, Bengio y Courville (2017), estas redes tienen como objetivo aproximar una función $y=f^*(x)$. En el caso de un clasificador, la función $y=f^*(x)$ mapea una entrada x a una categoría y. Una red feedforward define un mapeo $y=f(x;\theta)$ donde θ representa los parámetros que se aprenden para obtener la mejor aproximación de la función.
El término "feedforward" se refiere a cómo la información fluye a través de la red, desde la entrada x, pasando por los cálculos intermedios que definen f, hasta llegar a la salida y. En este tipo de redes no hay conexiones de retroalimentación donde las salidas del modelo se retroalimentan dentro de sí mismo.
La Figura 2, muestra cómo se compone una red neuronal feedforward. La red simple es una secuencia básica de capas de entrada, ocultas y de salida, mientras que la red profunda ilustra cómo múltiples capas ocultas aumentan la capacidad del modelo para aprender representaciones complejas de los datos. Esta estructura jerárquica facilita la captura de características significativas en los datos de entrada, logrando un rendimiento más alto en las aplicaciones de aprendizaje automático.

Figura 2.Redes feedforward simples y profundas



Las redes feedforward están típicamente representadas por la composición de muchas funciones diferentes, asociadas con un grafo acíclico dirigido que describe cómo se componen las funciones. El modelo incluye capas ocultas cuya dimensionalidad determina la

anchura del modelo. Cada capa oculta está valorada por vectores, y cada elemento del vector puede interpretarse como jugando un papel análogo a una neurona. Cada unidad recibe entrada de muchas otras unidades y computa su propio valor de activación.

Una función de activación utilizada en estas redes es LeakyReLU. LeakyReLU permite el aprendizaje de representaciones no lineales de los datos. A diferencia de la función ReLU (Rectified Linear Unit, por sus siglas en inglés) estándar, LeakyReLU permite un pequeño gradiente cuando la unidad no está activa, lo cual ayuda a mitigar el problema del ""dying ReLU" donde las unidades quedan permanentemente inactivas durante el entrenamiento.

Además, R. Guidotti et al. 2018, define una Red Neuronal Profunda (DNN), como un tipo de red neuronal que extiende las capacidades de las redes feedforward mediante el uso de múltiples capas ocultas para modelar relaciones no lineales complejas. La arquitectura de una DNN se compone de modelos estratificados, combinando unidades básicas para facilitar el flujo de datos desde la capa de entrada hasta la capa de salida sin retroalimentación.

2.1.3 Definición de Aprendizaje

Mitchell (1997), define que un programa de computadora aprende de la experiencia con respecto a una clase de tareas y una medida de rendimiento, si su rendimiento en esas tareas mejora con la experiencia.

2.1.3.1 Explorando la distinción entre Aprendizaje Supervisado y No Supervisado

El campo del aprendizaje automático se caracteriza por una diversidad de algoritmos y enfoques que se utilizan para extraer conocimiento de los datos. Entre estos enfoques, dos de los más fundamentales son el aprendizaje supervisado y el aprendizaje no supervisado. La distinción entre estos dos tipos de aprendizaje radica en la naturaleza de la experiencia que reciben los algoritmos durante el proceso de aprendizaje.

2.1.3.2 Aprendizaje Supervisado: etiquetando el camino hacia la predicción

En el aprendizaje supervisado, los algoritmos se enfrentan a conjuntos de datos que incluyen tanto características como etiquetas o valores objetivos. Por ejemplo, al entrenar un modelo para el reconocimiento de objetos en imágenes, cada imagen en el conjunto de datos vendría acompañada de una etiqueta que indica qué objeto se representa en la imagen. El objetivo del algoritmo es aprender a asociar características específicas con las etiquetas correspondientes, de modo que pueda hacer predicciones precisas sobre nuevas instancias basadas en estas asociaciones aprendidas.

2.1.3.3 Aprendizaje No Supervisado: descubriendo estructuras ocultas

En contraste, el aprendizaje no supervisado se enfrenta a conjuntos de datos que contienen solo características, sin etiquetas asociadas. Aquí, el objetivo principal es descubrir patrones o estructuras intrínsecas en los datos sin la ayuda de guías externas. Por ejemplo, los algoritmos de aprendizaje no supervisado pueden agrupar los datos en clusters de ejemplos similares o modelar la distribución de probabilidad de los datos, todo ello sin referencia a etiquetas predefinidas.

2.1.3.4 Capacidad, sobreajuste y subajuste en Modelos de Inteligencia Artificial

En el aprendizaje automático, la capacidad se refiere a la habilidad del modelo para adaptarse a una variedad de funciones. Vapnik y Chervonenkis (1971) establecieron los fundamentos teóricos de este concepto.

El sobreajuste se produce cuando un modelo memoriza el conjunto de datos de entrenamiento debido a una capacidad excesiva, en lugar de aprender patrones generales que puedan generalizarse a nuevos datos. Blumer et al. (1989) discutieron este concepto por primera vez en el contexto de la teoría del aprendizaje estadístico.

Por otro lado, el subajuste ocurre cuando un modelo no puede ajustarse adecuadamente a los datos de entrenamiento ni generalizar a datos no vistos debido a una capacidad insuficiente. Esta noción se originó en los primeros trabajos de Vapnik (1982) sobre la estimación de dependencias basadas en datos empíricos.

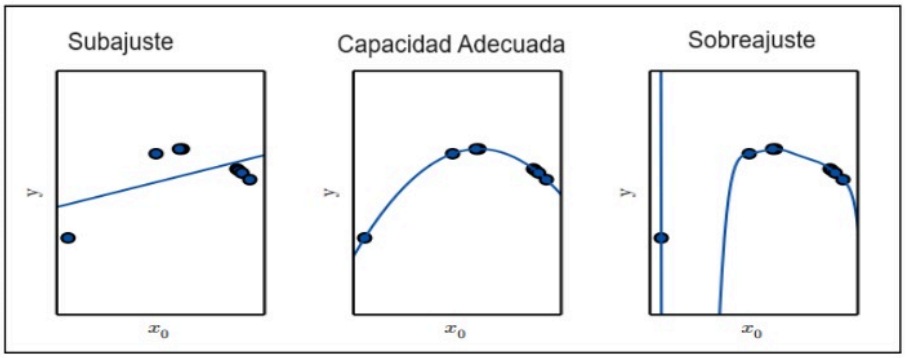
La Figura 3, ilustra tres modelos ajustados a un conjunto de datos de entrenamiento. Este conjunto de datos fue creado artificialmente mediante el muestreo aleatorio de valores x e y la determinación de y evaluando una función cuadrática (Goodfellow, Bengio y Courville, 2017).

En el primer panel, se observa el ajuste de una función lineal a los datos. Este modelo exhibe subajuste ya que no logra capturar la curvatura inherente en los datos, lo que indica que la relación entre x e y no está siendo correctamente representada debido a la simplicidad del modelo (Goodfellow, Bengio y Courville , 2017).

El segundo panel muestra el ajuste de una función cuadrática, la cual se ajusta adecuadamente a los datos de entrenamiento y generaliza bien a puntos no observados. Esto sugiere que el modelo no sufre de subajuste ni de sobreajuste significativo, ya que logra capturar la forma cuadrática de los datos sin complicaciones innecesarias (Goodfellow, Bengio y Courville, 2017).

En el tercer panel, se presenta el ajuste de un polinomio de grado 9 a los datos. Aquí se evidencia un claro caso de sobreajuste, donde el modelo es demasiado complejo para la estructura subyacente de los datos. Aunque este modelo pasa exactamente por todos los puntos de entrenamiento, introduce estructuras espurias como un valle profundo entre dos puntos de entrenamiento que no están presentes en la función verdadera. Esto demuestra que el modelo está memorizando el ruido en los datos en lugar de aprender patrones generales, lo que resulta en un rendimiento deficiente cuando se enfrenta a datos no observados (Goodfellow, Bengio y Courville, 2017).

Figura 3. Diferentes modelos se ajustan a conjuntos de datos de entrenamiento. En este contexto, se comparan tres modelos diferentes: uno lineal, otro cuadrático y un tercero polinómico de grado 9. Estos modelos se aplican a un problema donde la función subyacente verdadera es cuadrática (Goodfellow, Bengio y Courville, 2017).



2.1.3.5 Compensación entre sesgo y varianza para minimizar el Error Cuadrático Medio

Sesgo y Varianza:

- Sesgo: Mide la desviación esperada del estimador respecto al valor verdadero de la función o parámetro que estamos estimando. Es decir, cuánto se aleja, en promedio, nuestra estimación del valor verdadero.
- Varianza: Mide cuánto varía el estimador debido a la variación en la muestra de datos. En otras palabras, cuánto se dispersan las estimaciones del estimador alrededor de su valor esperado.

Elección entre Estimadores:

- A veces debemos elegir entre dos estimadores: uno con más sesgo y otro con más varianza.

- Para tomar esta decisión, podemos usar la validación cruzada, que empíricamente funciona muy bien en muchas tareas del mundo real.

Error Cuadrático Medio (MSE):

- El MSE es una medida del desvío total esperado (en términos de error cuadrático) entre el estimador y el valor verdadero del parámetro. Se calcula como la media de las diferencias al cuadrado entre las predicciones del modelo y los valores reales para cada punto de datos en el conjunto de entrenamiento. La fórmula del MSE es:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$$

Donde:

- m es el número de ejemplos en el conjunto de datos de entrenamiento.
- $y^{(i)}$ es el valor real del objetivo para el ejemplo i.
- $\hat{y}^{(i)}$ es la predicción del modelo para el ejemplo i.

El MSE incorpora tanto el sesgo como la varianza. Un estimador deseable es aquel con un MSE pequeño, lo cual significa mantener tanto el sesgo como la varianza bajo control.

Capacidad, Sobreajuste y Subajuste:

- La relación entre sesgo y varianza está vinculada a los conceptos de capacidad, sobreajuste (overfitting) y subajuste (underfitting).
- Incrementar la capacidad de un modelo tiende a incrementar la varianza y disminuir el sesgo.
- Existe una "curva en U" para el error de generalización como función de la capacidad, con una capacidad óptima que equilibra el sesgo y la varianza.

Consistencia:

- La consistencia de un estimador se refiere a su comportamiento a medida que aumenta el tamaño del conjunto de datos.

2.1.4 Visión por Computadora

La visión por computadora es una de las áreas más activas en investigación para aplicaciones de aprendizaje profundo, debido a que la visión es una tarea fácil para los humanos pero desafiante para las computadoras (Ballard et al., 1983). Muchas de las tareas estándar para los algoritmos de aprendizaje profundo incluyen el reconocimiento de objetos y el reconocimiento óptico de caracteres.

Este campo abarca una amplia gama de técnicas para procesar imágenes y una diversidad de aplicaciones, desde emular habilidades visuales humanas como el reconocimiento facial, hasta desarrollar nuevas capacidades visuales. Por ejemplo, detectar ondas sonoras a partir de las vibraciones generadas por objetos visibles en vídeos (Davis et al., 2014).

La mayoría de los avances en aprendizaje profundo para visión por computadora se centran en el reconocimiento y detección de objetos, como identificar qué objetos están presentes en una imagen, delimitar cada objeto con cuadros, transcribir símbolos desde una secuencia de imagen o etiquetar píxeles con la identidad del objeto.

Además del reconocimiento, la síntesis de imágenes mediante modelos profundos ha sido fundamental. Aunque la generación de imágenes desde cero no es típicamente considerada visión por computadora, estos modelos son importantes para la restauración de imágenes, reparando defectos o eliminando objetos no deseados.

Esta combinación de reconocimiento y síntesis de imágenes refleja cómo el aprendizaje profundo continúa transformando y expandiendo las capacidades de la visión por computadora.

2.1.5 Preprocesamiento

En muchas aplicaciones, es necesario realizar preprocesamiento porque la entrada original puede estar en una forma difícil de manejar para muchas arquitecturas de aprendizaje profundo. En visión por computadora, generalmente se requiere relativamente poco preprocesamiento. Las imágenes deben estandarizarse para que todos sus píxeles estén en el mismo rango, como [0,1] o [-1,1].

Combinar imágenes en [0,1] con imágenes en [0, 255] a menudo conduce a errores. El formateo para que las imágenes tengan la misma escala es el único tipo de preprocesamiento estrictamente necesario. Muchas arquitecturas de visión por computadora requieren imágenes de tamaño estándar, por lo que las imágenes deben recortarse o escalar para ajustarse a ese tamaño. Aunque este reescalado no siempre es estrictamente necesario.

Algunos modelos convolucionales pueden aceptar entradas de tamaño variable y ajustar dinámicamente el tamaño de sus regiones de agrupamiento para mantener constante el tamaño de salida (Waibel et al., 1989). Otros modelos convolucionales tienen salidas de

tamaño variable que se ajustan automáticamente al tamaño de la entrada, como los modelos para denoising o la etiquetación de cada píxel en una imagen (Hadsell et al., 2007).

El aumento de datos puede considerarse como una forma de preprocesamiento exclusiva del conjunto de entrenamiento, además de ser una estrategia para mejorar la generalización del modelo. Otra estrategia relacionada aplicable durante las pruebas es presentar al modelo diversas versiones de la misma entrada (por ejemplo, la misma imagen recortada en ubicaciones ligeramente diferentes) y dejar que las diferentes instancias del modelo voten para determinar la salida. Esta técnica puede entenderse como un enfoque de ensamblaje, útil para reducir el error de generalización.

Otros tipos de preprocesamiento se aplican tanto al conjunto de entrenamiento como al de prueba para estandarizar cada ejemplo y reducir la variabilidad que el modelo debe considerar. Reducir esta variabilidad puede disminuir tanto el error de generalización como el tamaño del modelo necesario para ajustar el conjunto de entrenamiento. Las tareas más simples pueden resolverse con modelos más pequeños, y estas soluciones simples suelen generalizar mejor.

Este tipo de preprocesamiento a menudo busca eliminar variabilidades en los datos de entrada que, aunque fáciles de describir para un diseñador humano, se considera que no son relevantes para la tarea en cuestión. Entrenando con grandes conjuntos de datos y modelos robustos, este tipo de preprocesamiento puede volverse innecesario, permitiendo al modelo aprender qué variabilidades son importantes por sí mismo. Por ejemplo, el sistema AlexNet para clasificar ImageNet solo realiza un preprocesamiento: resta la media de cada píxel basándose en los ejemplos de entrenamiento (Krizhevsky et al., 2012).

2.1.6 Procesamiento de Lenguaje Natural

El procesamiento del lenguaje natural (NLP) implica que las computadoras utilicen idiomas humanos, como el inglés o el francés. Los programas informáticos suelen trabajar con lenguajes especializados que permiten un análisis claro y eficiente. A menudo, los lenguajes naturales presentan ambigüedades que desafían una descripción formal. Entre las aplicaciones de NLP se encuentra la traducción automática, donde el sistema debe interpretar una oración en un idioma y producir una equivalente en otro. Muchas de estas aplicaciones se basan en modelos de lenguaje que establecen una distribución de probabilidad sobre secuencias de palabras, caracteres o bytes en un idioma natural.

2.2 Evaluación y Métricas

2.2.1 Métricas de Rendimiento

La elección de la métrica de error adecuada es esencial ya que guía todas las acciones futuras. También es importante definir el nivel de rendimiento deseado. En la mayoría de las aplicaciones, es imposible alcanzar un error absoluto de cero. El *error de Bayes* define la tasa mínima de error que se puede esperar, incluso con datos infinitos y la verdadera distribución de probabilidad, debido a la posible falta de información completa en las características de entrada y a la naturaleza estocástica del sistema. Además, la cantidad de datos de entrenamiento es finita, lo que también limita el rendimiento.

2.2.2 Métricas comunes y avanzadas

Eficiencia y tasa de error: Es común medir la precisión o la tasa de error del sistema. Sin embargo, muchas aplicaciones requieren métricas más avanzadas debido a la naturaleza de los errores y sus costos asociados.
Costo total: En situaciones donde un tipo de error es más costoso que otro, como en la detección de spam, se puede preferir medir una forma de costo total en lugar de la tasa de error. Por ejemplo, bloquear mensajes legítimos puede ser más costoso que permitir que pase spam.
Eficiencia y recall: Para clasificadores binarios destinados a detectar eventos raros, como una enfermedad rara, la precisión (fracción de detecciones correctas) y el recall (fracción de eventos verdaderos detectados) son métricas más adecuadas. Un clasificador que siempre dice que nadie tiene la enfermedad tendría precisión perfecta pero recall cero. Un clasificador que dice que todos tienen la enfermedad tendría recall perfecto pero baja precisión. Es común trazar una curva PR (Precisión-Recall) y usar la puntuación F para resumir el rendimiento:

$$F = \frac{2pr}{p+r}$$

Otra opción es reportar el área total bajo la curva PR.
Cobertura: En aplicaciones donde el sistema puede negarse a tomar una decisión, como en la transcripción de direcciones en Street View, la cobertura (fracción de ejemplos para los cuales el sistema puede producir una respuesta) es una métrica relevante. Se puede intercambiar cobertura por precisión. El objetivo para la tarea de Street View era alcanzar una precisión a nivel humano (98%) manteniendo una cobertura del 95%.
*Error Cuadrático Medio (MSE)***:** El Error Cuadrático Medio mide la media de los cuadrados de los errores, es decir, la diferencia promedio al cuadrado entre los valores predichos y los valores reales. Un MSE bajo indica que el modelo tiene una alta precisión en sus predicciones.
*Área Bajo la Curva (AUC)***:** El Área Bajo la Curva (AUC) de la Curva ROC es una métrica utilizada para medir la capacidad del modelo para distinguir entre clases positivas y negativas. Un AUC cercano a 1 indica un excelente rendimiento del modelo en la discriminación de clases.

2.2.3 Otras métricas posibles

Existen diversas métricas específicas a la aplicación, como las tasas de clics o las encuestas de satisfacción del usuario, entre otras. Es importante definir previamente qué métrica de rendimiento se va a optimizar y enfocarse en mejorar esa métrica para evaluar adecuadamente el progreso del sistema de aprendizaje automático.

Optimal Action Conformance (OAC%): La conformidad con la acción óptima (OAC%) mide en qué medida las acciones realizadas por el usuario coinciden con las acciones óptimas sugeridas por el sistema de soporte de decisión inteligente (IDS). Se calcula como el porcentaje de acciones realizadas que son consideradas óptimas según algún criterio predefinido.

Suboptimal Action Avoidance (SAA%): La capacidad de evitar acciones subóptimas (SAA%) indica qué tan bien los usuarios pueden identificar y evitar acciones que son subóptimas según las recomendaciones del sistema IDS. Se calcula como el porcentaje de acciones evitadas que son consideradas subóptimas en comparación con todas las acciones sugeridas por el sistema.

Perceived Preference (Pref%): La preferencia percibida (Pref%) refleja la preferencia subjetiva de los usuarios por un tipo específico de explicación o recomendación proporcionada por el sistema IDS. Esta métrica se obtiene a través de encuestas o evaluaciones donde los usuarios expresan su preferencia por diferentes tipos de explicaciones o recomendaciones.

Matriz de Confusión: La matriz de confusión proporciona una visión detallada del desempeño del modelo en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Ayuda a entender los tipos de errores que el modelo está cometiendo y su precisión en la predicción de cada clase.

Curvas de Pérdida y Precisión: Las curvas de pérdida y precisión son herramientas visuales que muestran cómo la precisión y la pérdida del modelo cambian con el tiempo durante el entrenamiento. Estas curvas son útiles para diagnosticar problemas como el sobreajuste y el subajuste.

2.3 Modelos Básicos

2.3.1 Modelos Básicos por Defecto

Modelos Simples: Para problemas sencillos y que pueden resolverse ajustando unos pocos pesos lineales, se recomienda el modelo de regresión logística.

Problemas Complejos ("AI-complete"): Si el problema pertenece a categorías como reconocimiento de objetos, reconocimiento de voz, traducción automática, etc., es mejor utilizar modelos de aprendizaje profundo.

2.3.2 Elección del Modelo según la Estructura de los Datos

Vectores de tamaño fijo (aprendizaje supervisado): Utilizar una red de alimentación directa con capas completamente conectadas.
Estructura topológica conocida (imágenes): Emplear una red convolucional.
Secuencias (entradas o salidas): Usar una red recurrente con puertas LSTMs (LSTMs - Long Short-Term Memory: Es una arquitectura de red neuronal recurrente (RNN) diseñada para modelar secuencias de datos. Las LSTMs son especialmente útiles cuando se necesita capturar relaciones a largo plazo en los datos. Tienen una "celda de memoria" que permite retener información. Esto las hace adecuadas para tareas como traducción automática, procesamiento del lenguaje natural y predicción de series temporales). GRU (Gated Recurrent Unit): Es otra variante de las redes neuronales recurrentes. Al igual que las LSTMs, las GRUs también están diseñadas para modelar secuencias. Sin embargo, las GRUs son más compactas y tienen menos parámetros que las LSTMs. Utilizan una "puerta de actualización" y una "puerta de reinicio" para controlar el flujo de información sin necesidad de una unidad de memoria de celda separada. Esto las hace más eficientes y a menudo funcionan bien en tareas similares a las LSTMs .
En estos casos, se recomienda comenzar con unidades lineales por tramos (ReLUs o sus generalizaciones como Leaky ReLUs, PreLus o Maxout).

2.3.3 Algoritmos de Optimización

SGD (Stochastic Gradient Descent): Es un algoritmo de optimización utilizado en el entrenamiento de modelos de aprendizaje automático. Funciona actualizando iterativamente los pesos del modelo basándose en el gradiente de la función de pérdida calculado con respecto a lotes pequeños de datos, en lugar de utilizar todo el conjunto de entrenamiento simultáneamente. Esta versión 'estocástica' acelera el proceso de entrenamiento y ayuda a evitar mínimos locales indeseados.

Adam: Es otra alternativa de optimización posible.

2.3.4 Regularización y Técnicas Adicionales

Normalización por lotes: Tiene efecto en el rendimiento de la optimización, especialmente para redes convolucionales y redes neuronales con funciones de activación sigmoidales. Si bien es razonable omitir la normalización por lotes en el primer intento, debería introducirse si la optimización resulta problemática.

Regularización suave: Debería incluirse desde el principio, a menos que el conjunto de entrenamiento contenga decenas de millones de ejemplos.

Dropout: Es un excelente regularizador, fácil de implementar y compatible con muchos modelos y algoritmos de entrenamiento.

2.4 Interpretabilidad en Modelos de Inteligencia Artificial

Guidotti (2018) definió la interpretabilidad como la capacidad de un modelo para explicar sus decisiones de manera que los usuarios puedan entender y confiar en ellas. Un modelo interpretable puede ser tanto global, donde se puede entender toda la lógica del modelo, como local, donde solo se pueden entender las razones de una decisión específica. Además, la interpretabilidad está relacionada con la precisión y la fidelidad del modelo, así como con aspectos éticos como la equidad y la privacidad.

Schwalbe y Finzel (2023) definieron la interpretabilidad como la capacidad de que las decisiones de un modelo de IA puedan ser explicadas global o localmente y que el propósito del modelo pueda ser entendido por un actor humano.

Un aspecto clave a considerar sobre la interpretabilidad es entender por qué un sistema, servicio o método debe ser interpretable. En algunos casos, una explicación no es necesaria si no se deben tomar decisiones basadas en el resultado de la predicción. Por ejemplo, si solo queremos saber si una imagen contiene un gato y esta información no afecta ninguna decisión importante ni tiene consecuencias significativas, no se requiere un modelo interpretable y se puede aceptar el uso de una "caja negra".

2.4.1 Beneficios de la Interpretabilidad

La interpretabilidad en los modelos de aprendizaje automático (ML) ofrece varios beneficios que van más allá de la simple precisión predictiva. A continuación, se detallan algunos de los principales beneficios:

- Confianza del Usuario: La interpretabilidad puede ayudar a desarrollar la confianza en los sistemas de ML. Lipton (2018) descompone la confianza en saber "con qué frecuencia un modelo es correcto" y "para qué ejemplos es correcto".
- Auditoría y Cumplimiento de Normativas: Es útil para auditar sistemas de ML y confirmar criterios adicionales más allá del rendimiento predictivo (Carvalho et al., 2019). Además, la interpretabilidad es coherente con la Regulación General de Protección de Datos (GDPR) de la UE, que otorga a los individuos el derecho a una explicación de las decisiones algorítmicas.

- Análisis Exploratorio y Descubrimiento Científico: Puede ser instrumental en el análisis exploratorio de datos y el descubrimiento científico. Por ejemplo, modelos de ML han sido utilizados para descubrir nuevas leyes físicas en ciencia de materiales y química cuántica (Liu et al., 2021; Schütt et al., 2019).
- Robustez y Transferibilidad: Los modelos interpretables pueden ayudar a diseñar modelos resistentes a entradas ruidosas y cambios en el dominio. Por ejemplo, los modelos aditivos generalizados se han utilizado para predecir riesgos de neumonía y corregir errores de confusión en los datos (Caruana et al., 2015).
- Equidad y Privacidad: En contextos donde los algoritmos de ML se utilizan para tomar decisiones sociales, económicas o médicas, es crucial evaluar su equidad y privacidad. Las explicaciones pueden ayudar a identificar disparidades demográficas y dependencias en información sensible (Carvalho et al., 2019).

2.4.2 Modelos Interpretables

Schwalbe y Finzel (2023), definieron los modelos interpretables como técnicas de aprendizaje automático que permiten rastrear relaciones causales o aprender representaciones estructuradas y que no requieren métodos adicionales para ser explicadas. La interpretabilidad del modelo se refiere al nivel de comprensión que se puede obtener del modelo utilizado para resolver la tarea. Puede ser intrínseca, donde el modelo en sí mismo es interpretable, o combinada, donde se integran modelos interpretables y no interpretables. También existen modelos auto-explicativos que proporcionan salidas adicionales que explican la predicción de manera automática.

Ejemplos:

- Modelos Intrínsecamente Interpretables: Incluyen árboles de decisión, redes bayesianas, modelos lineales, entre otros.

- Modelos Combinados: Fusionan modelos interpretables con modelos no interpretables, como redes neuronales y lógica difusa.

- Modelos Auto-explicativos: Proporcionan salidas adicionales que explican la predicción, como mapas de atención, representaciones desentrañadas y explicaciones textuales o multimodales.

En la literatura, se identifica un pequeño conjunto de modelos interpretables: árboles de decisión, reglas y modelos lineales(Freitas, 2014),(Huysmans et al., 2011),(Ribeiro et al., 2016)). Estos modelos se consideran fácilmente comprensibles e interpretables para los humanos.

2.4.3 Aprendizaje Automático Interpretable (Interpretable Machine Learning, iML)

Schwalbe y Finzel (2023) definieron el aprendizaje automático interpretable como el área de investigación que se enfoca en la creación de sistemas de AI interpretables.

2.5 Explicabilidad en Modelos de Inteligencia Artificial

En el campo de la inteligencia artificial, la explicabilidad se ha convertido en un aspecto determinante para garantizar la confianza y la aceptación de los sistemas por parte de los usuarios y la sociedad en general. A medida que la IA se aplica en una variedad de áreas importantes, desde la aprobación de préstamos hasta la asignación de recursos policiales y decisiones judiciales, surge la necesidad de comprender cómo estos sistemas llegan a sus conclusiones. En este sentido, la explicabilidad no solo implica entender el funcionamiento interno de los algoritmos, sino también abordar preocupaciones éticas relacionadas con la equidad, el sesgo y la responsabilidad en la toma de decisiones automatizada.

Aunque no hay una definición universalmente aceptada, diversos investigadores han aportado una perspectiva sobre el tema que a continuación se detalla:

2.5.1 Definición de Explicabilidad

La explicabilidad se refiere a la capacidad de los modelos de inteligencia artificial para justificar sus decisiones de una manera comprensible para los seres humanos. Esta capacidad es esencial para generar confianza entre los usuarios y las partes interesadas, especialmente en situaciones críticas.

Kass y Finin (1988) establecieron criterios claves para evaluar la calidad de una explicación, destacando la relevancia, la persuasión y la comprensibilidad como elementos esenciales.

Fischer et al. (1990) propusieron la idea de proporcionar explicaciones breves y minimalistas, con la flexibilidad de ampliarlas según las necesidades del usuario.

Gkatzia et al. (2016) sugirieron que la generación de lenguaje natural puede mejorar la toma de decisiones humanas, particularmente en entornos con datos inciertos. Subrayaron la utilidad de las explicaciones para aumentar la comprensión y la confianza en los modelos.

Biran y McKeown (2017) señalaron que la confianza en los modelos de aprendizaje automático depende de su capacidad para justificar sus decisiones, enfatizando la importancia de las explicaciones para establecer dicha confianza.

Doshi-Vélez y Kim (2017) argumentaron que una explicación efectiva actúa como una interfaz entre los humanos y el proceso de toma de decisiones del modelo, proporcionando una representación precisa y comprensible para los usuarios.

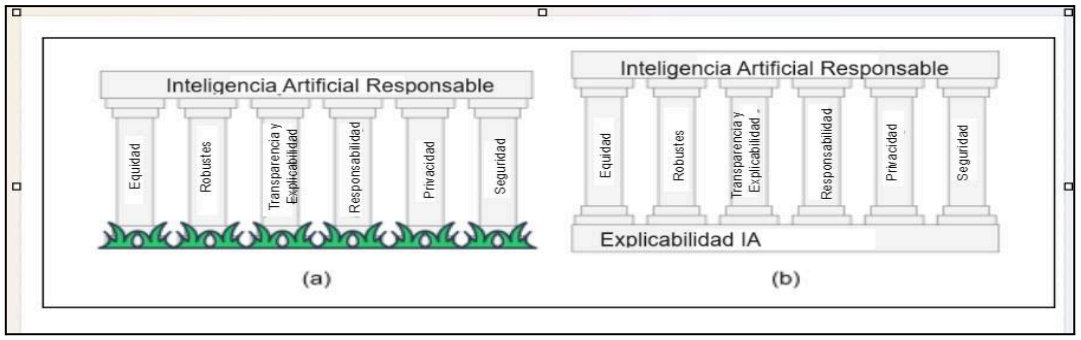
Miller (2019) definió una explicación como la respuesta a la pregunta "por qué", destacando la importancia de entender el razonamiento detrás de las decisiones del modelo.

Mueller et al. (2021) resaltaron la importancia no solo de generar explicaciones, sino también de evaluarlas mediante una medida formalizada.

Baker, S., y Xiang, W. (2023) (trabajo no revisado por pares) sostuvieron que la explicabilidad en la inteligencia artificial (XAI) se refiere a estrategias y procesos destinados a hacer comprensibles los modelos de AI para desarrolladores y usuarios finales, sin comprometer su rendimiento. Es esencial considerar a quién va dirigida la información, ya que las estrategias para entender un modelo pueden variar entre desarrolladores y usuarios no técnicos. Las motivaciones principales para la explicabilidad incluyen aumentar la transparencia y la confianza en los modelos de AI, especialmente en aplicaciones críticas.

Además, Baker y Xiang establecieron una relación entre la Explicabilidad en la Inteligencia Artificial (XAI) y la Inteligencia Artificial Responsable (RAI). Esto se ilustra en la Figura 4, donde se comparan dos enfoques de los marcos de RAI. En la Figura (a), se muestra el marco RAI según la definición común en la literatura, donde la explicabilidad se considera un pilar junto con la transparencia. En cambio, la Figura (b) representa el enfoque propuesto por Baker y Xiang, donde la explicabilidad se establece como la base para la responsabilidad. Este enfoque muestra que la XAI es fundamental para todos los pilares de responsabilidad.

Figura 4.Comparación de los marcos RAI(a) tal como se definen comúnmente en la literatura donde la explicabilidad es un pilar con la transparencia, y (b) como lo proponen Baker, S., & Xiang, W. (2023) con la explicabilidad como base para la responsabilidad.

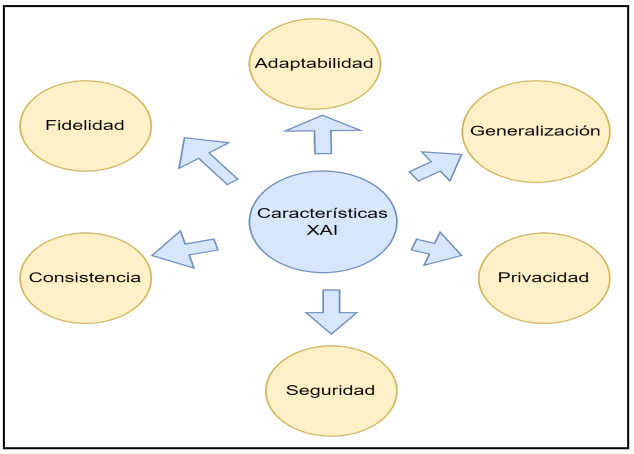


2.5.2 Inteligencia Artificial Explicable (Explainable Artificial Intelligence, XAI)

Schwalbe y Finzel (2023) definieron la inteligencia artificial explicable como el área de investigación centrada en explicar las decisiones de un sistema de IA.
Cada uno de los conceptos vistos previamente aborda la explicación de los modelos de inteligencia artificial de caja negra desde una perspectiva diferente, destacando aspectos como la generación de explicaciones claras, la comprensión del proceso de toma de decisiones del modelo y la accesibilidad de los resultados del modelo para los usuarios humanos.

En la Figura 5, se presentan las características esenciales de un sistema XAI. La figura incluye las siguientes palabras clave: fidelidad, adaptabilidad, consistencia, generalización, privacidad y seguridad. Estas características aseguran que el sistema de inteligencia artificial sea explicable y confiable en su funcionamiento.

Figura 5. Las características de un sistema XAI (elaboración propia)



2.5.2.1 Panorama de XAI

Los avances recientes en inteligencia artificial (IA) han ampliado sus aplicaciones en numerosos campos, pero también han incrementado la complejidad técnica de los modelos. Aunque se cree que los modelos de IA tienen el potencial de revolucionar diversos sectores, su implementación y aceptación en el mundo real requieren la confianza de profesionales, usuarios y otras partes interesadas (Dabiri et al., 2019). En la última década, el concepto de inteligencia artificial explicable (XAI) ha cobrado relevancia, reflejando los progresos en el campo de la IA. La XAI abarca herramientas y procesos diseñados para ayudar a los usuarios a entender tanto el funcionamiento interno de los modelos de IA como el mecanismo por el cual se generan los resultados.

Los modelos tradicionales de aprendizaje automático (ML), como la regresión lineal y los árboles de decisión, suelen ser menos complejos que los modelos de aprendizaje profundo (DL) y, por ende, más interpretables. Sin embargo, los modelos de DL generalmente ofrecen un rendimiento superior al de los modelos tradicionales. La dificultad para interpretar los resultados de los modelos de DL ha generado desconfianza y representa un obstáculo importante para su adopción práctica.

Gunning (2017) sostiene que las definiciones de "interpretable" y "explainable" pueden variar según el dominio y la tarea. En términos generales, la interpretabilidad se refiere a la capacidad de los humanos para comprender cómo un modelo ha llegado a una decisión. En contraste, la explicabilidad se relaciona con el entendimiento del mecanismo interno y la lógica del sistema de ML. Otros conceptos importantes son la confianza, que refleja el grado de certeza en la actuación del modelo, y la transferibilidad, que se refiere a la capacidad de los resultados del modelo para adaptarse a diferentes contextos. La XAI tiene como objetivo asegurar la confianza, la transferibilidad, la equidad, la accesibilidad y la interactividad.

En la Figura 6, se muestra una visión general de cómo los diferentes métodos y enfoques en aprendizaje automático pueden ser categorizados en función de su nivel de transparencia e interpretabilidad (Conard, DenAdel, & Crawford, 2023). A continuación, se describen los diferentes métodos ilustrados en la figura.

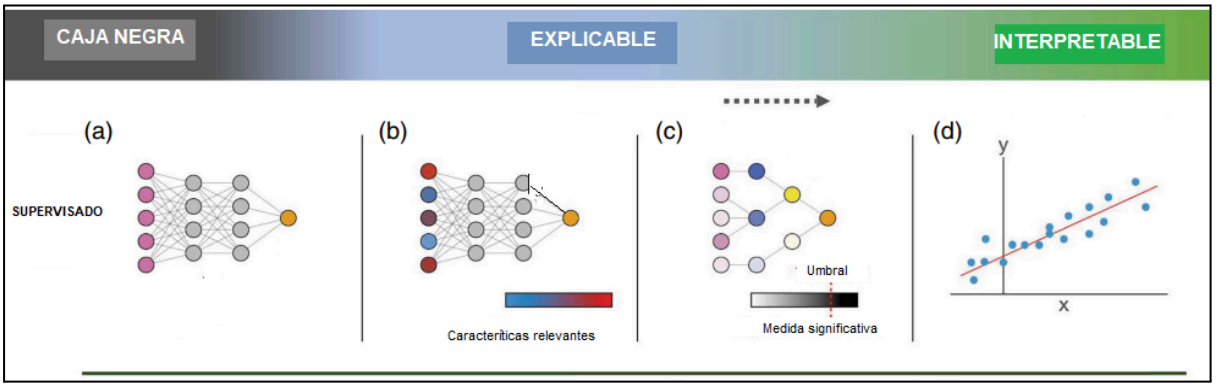
(a) **Red Neuronal Black Bo**x:Se muestra una arquitectura de red neuronal donde todos los neuronas de una capa están conectadas a todas las neuronas de la capa siguiente. Este tipo de modelo se llama "totalmente conectado". La desventaja es que se pierde la noción clásica del tamaño del efecto para cada característica, lo que dificulta la realización de pruebas de hipótesis estadísticas tradicionales.

(b) **Explicabilidad con Métodos Post Hoc**:Se refiere a la integración de métodos post hoc después del entrenamiento del modelo. Estos métodos permiten clasificar los inputs según su influencia en la salida (fenotipo), haciendo que el modelo sea explicable. Ejemplos de técnicas incluyen los métodos de atribución de características que explican el impacto de cada entrada en la predicción.

(c) **Métodos Supervisados Interpretables**:Tienen tres componentes clave: (i) un modelo probabilístico que los motiva, (ii) una medida del tamaño del efecto para cada característica de entrada, y (iii) una métrica de significancia para la selección de variables. Un enfoque intuitivo es usar redes parcialmente conectadas basadas en anotaciones biológicas o conocimientos científicos.

(d) **Modelos Lineale**s:Los modelos lineales son inherentemente interpretables porque permiten realizar pruebas de hipótesis bien controladas debido a sus componentes simples y transparentes.

Figura 6. Visión general de cómo los diferentes métodos y enfoques en aprendizaje automático pueden ser categorizados en función de su nivel de transparencia e interpretabilidad. Adaptado y modificado de Conard, DenAdel, & Crawford (2023).



2.5.2.2 Taxonomía de XAI

Los modelos interpretables se pueden clasificar en diferentes categorías. Una clasificación es si la interpretabilidad es intrínseca (ante hoc) o se establece después del entrenamiento (post hoc). Los modelos de ML/DL se consideran intrínsecamente interpretables si son autoexplicativos, como los árboles de decisión y la regresión logística. En contraste, la interpretabilidad post hoc requiere el uso de un segundo modelo para explicar el modelo original tras su entrenamiento. Un ejemplo es Grad-CAM, un método de XAI post hoc que visualiza el mapa de atención generando un mapa de calor en la capa convolucional más cercana a las capas completamente conectadas, lo cual ayuda a entender cómo el modelo llegó a su decisión (Dabiri et al., 2019).

Los métodos de XAI también pueden ser específicos del modelo o agnósticos. Un método específico del modelo está vinculado a la arquitectura y el funcionamiento interno de modelos de ML/DL concretos, mientras que un método agnóstico puede aplicarse a cualquier modelo o algoritmo, independientemente de su complejidad. SHAP (Shapley Additive Explanations) es un ejemplo de un método agnóstico.

Además, las explicaciones de XAI pueden ser locales o globales. Las explicaciones locales se refieren a una instancia específica, mientras que las explicaciones globales se aplican a todas las instancias del modelo. No hay un método superior en todos los casos; la elección del método adecuado depende del contexto específico.

2.5.2.3 Clasificación de los Métodos de XAI.

En el ámbito de la inteligencia artificial (IA), la capacidad de comprender y confiar en los modelos de aprendizaje automático es esencial, especialmente en contextos críticos donde las decisiones automatizadas pueden tener consecuencias significativas. La clasificación de los métodos de XAI según diferentes criterios(Vilone & Longo, 2021), como el escenario, el alcance y el tipo de problema, ofrece una guía estructurada para evaluar y seleccionar las herramientas más adecuadas para mejorar la explicabilidad de los modelos de inteligencia artificial.

Escenario:

> **Previamente (Ante-hoc):** Métodos que integran la explicabilidad en el modelo desde el inicio de su desarrollo. Estos modelos están diseñados para ser interpretables de manera inherente, es decir, su estructura y funcionamiento son transparentes y comprensibles sin necesidad de técnicas adicionales.
>
> **Posteriormente (Post-hoc)**:Explicaciones generadas después que el modelo hizo una predicción.Se utiliza un conjunto diferente de métodos para proporcionar las explicaciones necesarias sin necesidad de entender el modelo original. Esto implica la creación de un modelo secundario como sustituto
>
> - **Modelo agnóstico**:Métodos que pueden aplicarse a cualquier tipo de modelo de aprendizaje automático, sin depender de la estructura interna del modelo. Estos métodos tratan al modelo como una caja negra y se enfocan en las relaciones entre las entradas y las salidas.
> - **Modelo específico:**Métodos diseñados para explicar modelos específicos, aprovechando su estructura interna. Estos métodos son personalizados y optimizados para ciertos tipos de modelos, como redes neuronales, árboles de decisión, etc.

Alcance: Evalúa cómo las salidas del modelo se relacionan con las entradas.

- Global:Proporciona explicaciones generales para todo el modelo.(ejemplo: SHAP, que también puede ofrecer explicaciones locales).
- Local:Se centra en una sola instancia o caso (ejemplo: LIME)

Tipo de problema:

- **Clasificación**:.Métodos diseñados para explicar modelos que asignan instancias a categorías discretas

- **Regresión**:Métodos diseñados para explicar modelos que predicen valores continuos

Dato de entrada:

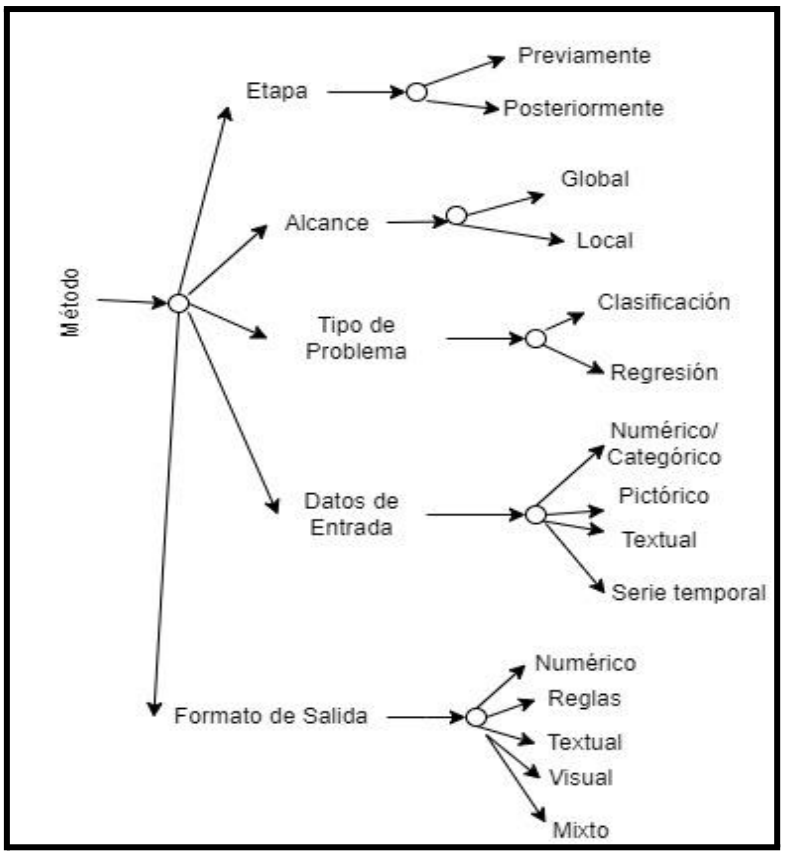
- **Numérico/Categórico**:Datos estructurados en forma de números o categorías discretas. Las explicaciones para estos datos suelen centrarse en la importancia y el impacto de diferentes características
- **Imágen**:Datos en formato visual. Las explicaciones para imágenes pueden incluir mapas de calor que muestran qué partes de la imagen influyeron en la predicción
- **Texto:**Datos en formato de texto. Las explicaciones para texto pueden resaltar palabras o frases clave que contribuyeron a la predicción
- **Serie temporal:** Datos que representan una secuencia de puntos en el tiempo. Las explicaciones pueden identificar patrones temporales y eventos críticos que influyen en las predicciones

Formato de salida:

- **Reglas:** Explicaciones basadas en reglas que proporcionan una representación estructurada y compacta de un conjunto de instrucciones lógicas.
- **Información Numérica:** Explicaciones que utilizan datos numéricos para ilustrar el razonamiento detrás de una decisión del modelo.
- **Información Textual:** Explicaciones que utilizan texto para describir por qué se ha llegado a una determinada conclusión o predicción.
- **Información Visual:** Explicaciones que emplean imágenes o visualizaciones para mostrar cómo el modelo ha llegado a su predicción.
- **Combinación de los Formatos Anteriores:** Explicaciones que combinan dos o más de los formatos mencionados anteriormente para proporcionar una comprensión más completa y accesible.Por ejemplo, una combinación de texto y visualizaciones puede ofrecer una explicación más rica y detallada.

La Figura 7 muestra la clasificación de los métodos de inteligencia artificial explicable según los distintos criterios. (Vilone, G.; Longo, L (2021)).

Figura 7. Clasificación de los métodos de XAI en un sistema jerárquico definido por Vilone, G.; Longo, L (2021).



Capítulo 3

Revisión de la Literatura

3.1 Introducción

La inteligencia artificial explicable (XAI) es fundamental en la investigación de IA, ya que busca desarrollar modelos y sistemas que puedan explicar de manera comprensible su comportamiento y procesos de toma de decisiones. Entender y explicar estas decisiones garantiza la aceptación, confianza y aplicabilidad de los modelos en diversos campos.

Este capítulo aborda la estrategia de búsqueda bibliográfica, la distribución de artículos recuperados de diversas bibliotecas digitales, y los criterios para la inclusión y exclusión de estudios. También se detalla la metodología aplicada a los artículos seleccionados y cómo se presentarán las conclusiones finales.

3.2 Búsqueda y selección de estudios

La siguiente tabla detalla las palabras clave utilizadas en la revisión sistemática, alineadas con el enfoque PICO propuesto por Kitchenham et al. (2008). Cada uno de los aspectos de la tabla, junto con sus respectivas palabras clave, proporciona un marco detallado para la búsqueda y análisis de la literatura relevante (véase Tabla 1).

Tabla 1 Enfoque Pico propuesto por Kitchenham et al.,2008.

Aspecto	Palabras claves
Población (Population)	("artificial intelligence" OR "machine learning" OR “deep learning”) AND (model OR system OR algorithm) / ("inteligencia artificial" O "aprendizaje automático" O “aprendizaje profundo”) Y (modelo O sistema O algoritmo)
Intervention (Intervención)	(explanation OR explainability OR interpretable) AND (taxonomy OR method OR technique) / (explicación O explicabilidad O interpretable) Y (taxonomía O método O técnica)
Comparison (Comparación)	(contrast OR differentiation) / (contraste O diferenciación)
Outcome (Resultado)	(advances OR proposals OR taxonomies) / (avances O propuestas O taxonomías)

La pregunta de investigación, *"¿Cuáles son los avances más recientes en la mejora de la explicabilidad de los modelos de machine learning considerados 'caja negra', en comparación con enfoques anteriores, y cuál es el impacto de dichas mejoras, en términos de propuestas, taxonomías y otros resultados relevantes?"*, sirvió como guía para seleccionar las palabras clave adecuadas en cada aspecto de la búsqueda.

3.3 Estrategia para la búsqueda bibliográfica

La búsqueda bibliográfica se realizó en bases de datos de investigación académica, incluyendo ACM Digital Library, SEDICI, IEEE Xplore y Google Scholar. Para llevar a cabo la búsqueda de manera sistemática, se combinaron las siguientes palabras clave, basadas en el enfoque PICO:
("artificial intelligence" OR "machine learning" OR "deep learning") AND (model OR system OR algorithm) AND (explanation OR explainability OR interpretable) AND (taxonomy OR method OR technique) AND (contrast OR differentiation) AND (advances OR proposals OR taxonomies)
La búsqueda se limitó a documentos publicados desde el 1 de enero de 2022 hasta el 5 de mayo de 2024 y se centró en los títulos y resúmenes de los artículos, priorizando aquellos de libre acceso.
Los resultados de esta búsqueda se presentan en el Anexo II, titulado *Papers Recuperados sobre Modelos de Explicabilidad en IA*. A partir de estos trabajos, se aplicaron los criterios de inclusión definidos, lo que llevó a la selección de los artículos detallados en el Anexo I, titulado *Análisis Detallado de 30 Trabajos Seleccionados sobre Avances Recientes en la Explicabilidad de Modelos de IA*. Estos son los documentos sobre los cuales se realizó el análisis.

3.4 Criterios de inclusión y exclusión

Para abordar de manera efectiva la pregunta de investigación planteada, es importante establecer criterios de inclusión y exclusión. Estos criterios aseguran que los artículos seleccionados proporcionan información sobre los avances recientes en la explicabilidad de los modelos de machine learning considerados 'caja negra', en comparación con enfoques anteriores, y que evalúen el impacto de dichas mejoras en términos de propuestas, taxonomías y otros resultados relevantes. A continuación, se detallan los criterios utilizados en esta revisión teniendo en cuenta la pregunta de investigación (ver punto 3.2):

3.4.1 Criterios de Inclusión

1. **Avances recientes en la explicabilidad de modelos 'Caja Negra':**

- Identificación de Nuevas Metodologías: El abstract menciona alguna técnica innovadora desarrollada recientemente para mejorar la explicabilidad de los modelos 'caja negra'.

2. **Comparación con enfoques anteriores:**
   - Comparación con Métodos Tradicionales: El abstract compara explícitamente las nuevas técnicas con métodos más antiguos, proporcionando una evaluación clara de las mejoras en términos de explicabilidad.
3. **Impacto de las mejoras:**
   - Propuestas y Taxonomías: El abstract menciona resultados específicos, propuestas nuevas, frameworks o taxonomías que indiquen el impacto de las mejoras en la explicabilidad.
   - Resultados Relevantes: Se presentan casos de uso prácticos, estudios de caso, o cualquier otro resultado que demuestre el impacto práctico de las mejoras en la explicabilidad.
4. **Fecha de publicación:** Desde el 1 de enero de 2022 hasta el 5 de mayo de 2024**.**

3.4.2 Criterios de Exclusión

1. **Falta de Enfoque en Explicabilidad:**
   - Explicabilidad No Evaluada: Excluir trabajos que no abordan directamente la explicabilidad o que solo mencionan la explicabilidad de manera superficial sin evaluarla.
2. **Modelos Transparentes:**
   - No Considerados 'Caja Negra**':** Excluir investigaciones centradas en modelos que son inherentemente interpretables (por ejemplo, regresión lineal, árboles de decisión simples) y no se consideran 'caja negra'.
3. **Falta de Comparación con Métodos Anteriores:**
   - Sin Comparación Explícita: Excluir estudios que no comparen explícitamente las nuevas técnicas con métodos anteriores en términos de explicabilidad.
4. **Sin Evaluación de Impacto:**
   - Impacto No Medido: Excluir investigaciones que no midan de alguna forma el impacto de las mejoras propuestas en términos de explicabilidad, utilidad práctica, o comprensión por parte de los usuarios.
5. **Paper sin libre acceso:** Excluir trabajos que no estén disponibles en acceso libre.

3.5 Metodología de Análisis de los Artículos Seleccionados

Los artículos fueron seleccionados siguiendo los criterios de inclusión y exclusión detallados en las secciones 3.4.1 y 3.4.2. Se realizó un análisis de los mismos para abordar la pregunta

de investigación planteada. El proceso de análisis se llevó a cabo en varias etapas, descritas a continuación:

3.5.1 Lectura y Comprensión de los Artículos

En esta etapa, se leyó detalladamente cada artículo seleccionado para comprender a fondo las nuevas metodologías propuestas, las comparaciones con enfoques anteriores y el impacto de las mejoras en términos de explicabilidad.

3.5.2 Categorización de las Técnicas de Explicabilidad

Se categorizó cada artículo según las técnicas de explicabilidad que presentan, permitiendo una organización sistemática de las metodologías analizadas.

3.5.3 Comparación de Técnicas con Enfoques Anteriores

Se realizó una comparación detallada de las nuevas técnicas con los enfoques anteriores. Esta comparación se centró en evaluar las mejoras en la explicabilidad que ofrece cada nueva técnica. Se prestó especial atención a cómo se han evaluado estas mejoras y a los resultados obtenidos.

3.5.4 Evaluación del Impacto de las Mejoras

Se analizaron los artículos para identificar las propuestas, taxonomías, frameworks y otros resultados específicos que indican el impacto de las mejoras en la explicabilidad. También se evaluaron casos de uso prácticos y estudios de caso presentados en los artículos para entender el impacto práctico de las nuevas metodologías.

3.5.5 Síntesis de resultados

Se sintetizaron los hallazgos de los análisis anteriores para proporcionar una visión general comprensiva de los avances recientes en la explicabilidad de los modelos 'caja negra'.

3.5.6 Elaboración de conclusiones

Finalmente, se elaboraron conclusiones basadas en los resultados del análisis, discutiendo las implicaciones de los avances en la explicabilidad y su relevancia para el campo.

Capítulo 4

Análisis de Metodologías y Aplicaciones XAI

4.1 Introducción al Análisis de XAI

En este capítulo se examinan los artículos que cumplen con los criterios de inclusión establecidos para esta revisión del estado del arte. Estos abordan aspectos importantes en el campo de la inteligencia artificial explicable (XAI), explorando metodologías y aplicaciones prácticas en diversos dominios. El objetivo principal es evaluar cómo estas investigaciones contribuyen al avance de la XAI, enfocándose en mejorar la transparencia, interpretabilidad y robustez de los modelos de aprendizaje automático.

4.1.1 Trabajos que cumplen con los criterios de inclusión definidos

A continuación se presenta la Tabla 2, que resume los artículos que cumplen con los criterios de inclusión definidos. Cada artículo ha sido evaluado en función de su contribución significativa al campo de la XAI, considerando aspectos como la interpretabilidad de los modelos, la robustez frente a perturbaciones y la aplicación práctica en diversos dominios. La tabla ofrece una visión general de los papers seleccionados, destacando sus enfoques y resultados principales.

Cada artículo listado en la tabla incluye un enlace al Anexo I, que proporciona una visión detallada de los enfoques, metodologías y resultados, ofreciendo una base para comprender cómo cada uno contribuye a la mejora de la explicabilidad en el contexto de la XAI.

Tabla 2. Trabajos que cumplen con los criterios de inclusión definidos (véase punto 3.4)

Nº	Título del paper	Cita	Comentarios
1	Equidad contrastiva contrafactual en la toma de decisiones algorítmica.	Mutlu, E. Ç., Yousefi, N., & Garibay, O. O. (2022). Contrastive counterfactual fairness in algorithmic decision-making. En Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22) (pp. 499–507). https://doi.org/10.1145/3514094.3534143	Mutlu, Yousefi y Garibay (2022) exploran la **equidad contrastiva contrafactual** en la toma de decisiones algorítmica en su trabajo presentado en la conferencia AAAI/ACM sobre AI, ética y sociedad. El estudio aborda el desafío de equilibrar la equidad y la precisión en los algoritmos de inteligencia artificial. Propone enfoques de **aprendizaje causal justo** que modelan relaciones de causa y efecto para identificar y mitigar sesgos en los sistemas de AI. Introduce un **criterio de equidad contrafactual contrastiva** y un **método de aumento de datos consciente de la equidad**, evaluado en los conjuntos de datos UCI Adult y German Credit. Este método, que es independiente del modelo, demuestra ser eficaz para reducir la inequidad en los sistemas de IA.
2	Explicabilidad federada para la caracterización de anomalías de red	Sáez-de-Cámara, X., Flores, J. L., Arellano, C., Urbieta, A., & Zurutuza, U. (2023). Federated Explainability for Network Anomaly Characterization. In *Proceedings of the 26th*	El documento presenta un enfoque para mejorar la explicabilidad en sistemas de detección de intrusiones basados en ML, especialmente en entornos distribuidos como el Internet de las cosas (IoT). Destaca la importancia de proporcionar información contextual para que los analistas de seguridad comprendan por qué se clasificó una muestra como anómala y cómo correlacionar diferentes tipos de anomalías. La metodología

		International Symposium on Research in Attacks, Intrusions and Defenses (RAID '23) (pp. 346–365). https://doi.org/10.1145/3607199.3607234	propuesta adapta algoritmos de explicabilidad, agrupamiento y validación de grupos para extraer patrones en muestras anómalas e identificar amenazas en toda la red. Los resultados demuestran la utilidad de este enfoque en conjuntos de datos de detección de intrusiones del mundo real.
3	RoCourseNet: Entrenamiento robusto de un modelo de recurso consciente de predicción	Guo, H., Jia, F., Chen, J., Squicciarini, A., & Yadav, A. (2023). RoCourseNet: Robust Training of a Prediction Aware Recourse Model. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (CIKM '23), 619–628. https://doi.org/10.1145/3583780.3615040	Guo et al. (2023) presentan **RoCourseNet** en la conferencia CIKM '23, un modelo diseñado para generar **recursos conscientes de la predicción** de manera robusta. El enfoque se basa en **explicaciones contrafácticas** (CF), que muestran qué modificaciones en las características de entrada podrían haber llevado a una predicción diferente del modelo de aprendizaje automático. RoCourseNet se enfoca en crear **recursos robustos**, que son válidos incluso si hay cambios en la distribución de los datos, mejorando así la estabilidad y fiabilidad de las explicaciones proporcionadas.
4	SAMCNet: Hacia un enfoque de inteligencia artificial explicable para clasificar datos oncológicos MxIF.	Farhadloo, M., Molnar, C., Luo, G., Li, Y., Shekhar, S., Maus, R. L., Markovic, S., Leontovich, A., & Moore, R. (2022). SAMCNet: Towards a Spatially Explainable AI Approach for Classifying MxIF Oncology Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD '22), (pp. 2860–2870). https://doi.org/10.1145/3534678.3539168	Farhadloo et al. (2022) presentan **SAMCNet** en la conferencia KDD '22, un enfoque de **inteligencia artificial explicable espacialmente** para la clasificación de datos oncológicos MxIF. Este enfoque se centra en la **explicabilidad espacial**, que busca comprender cómo la organización espacial de los datos contribuye a la clasificación entre dos grupos: **respondedores** (pacientes que muestran una respuesta positiva a un tratamiento) y **no respondedores** (pacientes que no muestran una respuesta significativa). SAMCNet se aplica a datos multiclase y tiene aplicaciones en la investigación biomédica para el desarrollo de nuevas terapias contra el cáncer y en la ecología microbiana. El método propuesto facilita la interpretación de cómo las características espaciales influyen en las decisiones del modelo, proporcionando una visión más clara de las relaciones en los datos.
5	Explicaciones basadas en submetas para sistemas de soporte de decisiones inteligentes no confiables.	Das, D., Kim, B., & Chernova, S. (2023). Subgoal-based explanations for unreliable intelligent decision support systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)* (pp. 240–250). https://doi.org/10.1145/3581641.35840	Das, Kim y Chernova (2023) presentan un enfoque innovador para mejorar la **interpretabilidad** de los **sistemas de soporte de decisiones inteligentes** (IDS) que a menudo pueden ser poco fiables o fallar en situaciones complejas. El artículo introduce **explicaciones basadas en submetas**, un tipo de explicación que proporciona información adicional sobre la **submeta** hacia la cual la acción recomendada por el IDS contribuye. Estas explicaciones ayudan a los usuarios a entender mejor el contexto de las recomendaciones y a distinguir entre opciones óptimas y subóptimas. El estudio demuestra que este enfoque no solo mejora el rendimiento de los usuarios al usar recomendaciones del IDS, sino que también les ayuda a manejar mejor las fallas del IDS. Además, las explicaciones basadas en submetas son preferidas por los usuarios y resultan ser útiles para entrenar a los usuarios en la tarea subyacente.
6	LCNN: Arquitectura Ligera de CNN para la Identificación de Características de	Begum, M., Alam, M., Islam, M. R., & Hossain, M. A. (2024). LCNN: Lightweight CNN Architecture for Software	Begum, Alam, Islam y Hossain (2024) presentan **LCNN**, una **arquitectura ligera de redes neuronales convolucionales (CNN)** diseñada para la identificación de características de defectos de software, utilizando técnicas de **inteligencia artificial explicativa (XAI)**. El

	Defectos de Software Utilizando Inteligencia Artificial Explicable.	Defect Feature Identification Using Explainable AI. *IEEE Access*, 12, 55744-55756. DOI 10.1109/ACCESS.2024.3388489 https://ieeexplore.ieee.org/document/10499820	artículo enfatiza la necesidad de mejorar la **transparencia** e **interpretabilidad** en los modelos de IA para esta tarea. Se exploran dos variantes de LCNN (1D-CNN y 2D-CNN) y técnicas de preprocesamiento de datos como **SMOTE** para manejar el desequilibrio en los datos. Se comparan los resultados de estas técnicas en términos de **precisión**, **error cuadrático medio (MSE)** y **área bajo la curva (AUC)**. Además, se evalúan métodos XAI como **LIME** y **SHAP** para explicar las características identificadas por el modelo. Los resultados muestran que el LCNN en su versión 2D supera al 1D-CNN en la identificación de defectos de software. LIME se destaca por su capacidad para visualizar de manera efectiva las características de los defectos, proporcionando una comprensión más clara de sus causas subyacentes
7	Deep Prototypical-Parts facilita la identificación morfológica de cálculos renales y es competitivamente robusto ante perturbaciones fotométricas.	Flores-Araiza, D., et al. (2023). Deep Prototypical-Parts Ease Morphological Kidney Stone Identification and are Competitively Robust to Photometric Perturbations. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 295-304). Vancouver, BC, Canada. doi: 10.1109/CVPRW59228.2023.00035 . https://ieeexplore.ieee.org/document/10208481	Flores-Araiza et al. (2023) presentan un enfoque de **aprendizaje profundo (DL)** para la **identificación de cálculos renales**, que prioriza la **explicabilidad** del modelo. Utilizan **Prototypical Parts (PPs)** para generar explicaciones interpretables sobre las decisiones del modelo. Aunque el enfoque basado en PPs muestra una **precisión promedio ligeramente inferior** en comparación con modelos DL no interpretables, destaca por su **robustez frente a perturbaciones fotométricas** en las imágenes. Este enfoque sugiere que la incorporación de PPs no solo facilita la interpretación de los resultados, sino que también mejora la **robustez** del modelo ante cambios en las condiciones de las imágenes.
8	Sistemas Explicativos Basados en Reglas Difusas para Redes Neuronales Profundas: Desde Explicabilidad Local hasta Comprensión Global.	Aghaeipoor, F., Sabokrou, M., & Fernández, A. (2023). Fuzzy Rule-Based Explainer Systems for Deep Neural Networks: From Local Explainability to Global Understanding. *IEEE Transactions on Fuzzy Systems*, 1–12. https://doi.org/10.1109/TFUZZ.2023.3243935	Aghaeipoor et al. (2023) presentan **sistemas explicativos basados en reglas difusas** para **redes neuronales profundas** (DNN), con el objetivo de mejorar la **explicabilidad local** y **comprensión global** de estos modelos complejos. La investigación destaca que, aunque la extracción de reglas post-hoc es útil para entender la lógica interna de las redes neuronales, las representaciones numéricas en las reglas pueden no ser intuitivas. Por ello, el artículo propone el uso de **conjuntos y reglas difusas** para ofrecer una representación más semántica y comprensible. El algoritmo desarrollado aprende un conjunto compacto y preciso de reglas difusas basadas en la **importancia de las características** (atribuciones de valores) extraídas de las redes neuronales entrenadas. Estas reglas no solo permiten una mejor **interpretación local** (entender cómo el modelo llega a decisiones específicas) sino también una **comprensión global** (entender el funcionamiento general del modelo). Los resultados evaluados en diversas aplicaciones demostraron que los explicadores difusos mantenían la **fidelidad** y **precisión** de las redes neuronales profundas originales, con la ventaja adicional de una **menor complejidad** y una mejor **comprensión**.
9	*Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy*	Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N.,	Ali et al. (2023) realizan una **revisión exhaustiva** sobre el estado actual de la **Inteligencia Artificial Explicable (XAI)** y su relevancia para mejorar la **comprensión y confianza** en los modelos de inteligencia artificial, que a menudo se perciben como cajas negras difíciles de interpretar. El artículo destaca la **necesidad urgente** de métodos XAI para

	Artificial Intelligence	& Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805	aumentar la confianza en los modelos de IA y ofrece una visión completa de las técnicas recientes en XAI, especialmente en el contexto del **aprendizaje supervisado**. La investigación clasifica **XAI** en cuatro categorías principales: 1. **Explicabilidad de Datos:** Cómo se presentan y explican los datos utilizados en el entrenamiento. 2. **Explicabilidad de Modelos:** Cómo se interpreta el funcionamiento interno de los modelos. 3. **Explicabilidad Post-Hoc:** Cómo se generan explicaciones después de que el modelo ha hecho una predicción. 4. **Evaluación de Explicaciones:** Cómo se mide la efectividad de las explicaciones proporcionadas. El artículo también presenta **métricas de evaluación**, **herramientas de código abierto**, y **conjuntos de datos disponibles** que facilitan la investigación en XAI. Se analizan 410 artículos clave publicados entre 2016 y 2022, proporcionando una **visión crítica** de los avances en XAI y **direcciones futuras** para la investigación. Además, se subraya la importancia de adaptar las explicaciones a diferentes tipos de usuarios para que sean útiles y comprensibles. Este artículo está dirigido a **investigadores de XAI** y a profesionales de otras disciplinas que buscan **métodos efectivos** para implementar y entender XAI en sus aplicaciones.
10	Explicaciones Lógicas Basadas en Entropía de Redes Neuronales.	Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., & Melacci, S. (2022). Entropy-Based Logic Explanations of Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 36(6), 6046–6054. https://doi.org/10.1609/aaai.v36i6.2055_1	**Barbiero et al. (2022)** proponen un enfoque para extraer **explicaciones lógicas** de **redes neuronales** utilizando **Lógica de Primer Orden** y un criterio basado en **entropía**. Este método se centra en identificar automáticamente los conceptos más relevantes dentro de los modelos de aprendizaje profundo. **Aspectos Clave del Artículo:** 1. **Método Propuesto:** ○ **Lógica de Primer Orden (FOL):** Se utiliza para formalizar las explicaciones lógicas de las decisiones tomadas por las redes neuronales. ○ **Criterio Basado en Entropía:** Este criterio ayuda a identificar los conceptos y características más importantes para la explicación, facilitando una comprensión más clara de cómo el modelo llega a sus decisiones. 2. **Estudios de Caso:** ○ Se presentan cuatro estudios de caso que aplican el enfoque en diferentes dominios críticos, desde **datos clínicos** hasta **visión por computadora**. ○ Los estudios demuestran que el enfoque basado en entropía permite extraer explicaciones lógicas **concisas** y **relevantes**, facilitando una mejor interpretación del

			modelo. 3. **Comparación con Modelos de Caja Blanca:** ○ El enfoque propuesto no solo mejora la **explicabilidad** de las redes neuronales, sino que también supera a los **modelos de caja blanca** en términos de **precisión de clasificación**. 4. **Aplicaciones y Beneficios:** ○ La capacidad de extraer explicaciones lógicas precisas y significativas tiene un impacto significativo en **dominios críticos para la seguridad**, como la medicina y la visión por computadora.
11	Redes Explicadas por Lógica.	Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., & Melacci, S. (2023). Logic Explained Networks. *Artificial Intelligence*, *314*, 103822. https://doi.org/10.1016/j.artint.2022.103822	Ciravegna et al. (2023) presentan un enfoque innovador para la **Inteligencia Artificial Explicable (XAI)** con el desarrollo de **Logic Explained Networks (LENs)**. Estas redes neuronales se diseñan para ser interpretables, proporcionando explicaciones claras y comprensibles en términos de **Lógica de Primer Orden (FOL)**. **Características principales de LENs:** 1. **Predicados Comprensibles:** LENs requieren que sus entradas sean predicados que los humanos puedan entender fácilmente. 2. **Explicaciones en FOL:** Las explicaciones generadas por LENs se expresan mediante fórmulas simples de FOL, facilitando la comprensión de cómo se toman las decisiones. 3. **Aplicaciones Amplias:** Los LENs pueden actuar tanto como **clasificadores con capacidad explicativa** como redes adicionales para **explicar clasificadores de caja negra**. **Ventajas de LENs:** ● **Superan a Modelos Tradicionales:** En comparación con modelos de caja blanca establecidos como árboles de decisión y listas de reglas bayesianas, los LENs pueden ofrecer **clasificaciones superiores** y **explicaciones más completas**. ● **Versatilidad en el Aprendizaje:** Los LENs son aplicables a **tareas de aprendizaje supervisado y no supervisado**, demostrando su flexibilidad y eficacia en diversos escenarios. **Resultados Experimentales:** Los experimentos realizados en diferentes conjuntos de datos y tareas confirmaron que los LENs no solo superan a los modelos tradicionales en precisión, sino que también proporcionan explicaciones más detalladas y significativas.
12	Aprendizaje Profundo con Restricciones Lógicas	Giunchiglia, E., Stoian, M. C., & Lukasiewicz, T. (2022). *Deep Learning with Logical Constraints* (arXiv:2205.00523). arXiv. http://arxiv.org/abs/2205.00523	**Giunchiglia et al. (2022)** exploran la integración de **conocimientos previos** expresados en **lógica de primer orden (FOL)** en modelos de **aprendizaje profundo** para mejorar tanto su **rendimiento** como su **explicabilidad**. **Aspectos Clave del Artículo:**

			1. **Integración de Conocimientos Previos:** Se presenta un enfoque para incorporar **restricciones lógicas** en redes neuronales profundas, utilizando FOL para formalizar y agregar conocimientos específicos al proceso de aprendizaje. 2. **Beneficios de la Integración:** ○ **Mejora del Rendimiento:** La integración de estas restricciones permite que el modelo aproveche el conocimiento experto, lo que puede conducir a un mejor ajuste y generalización en las tareas de aprendizaje. ○ **Mayor Explicabilidad:** Al incorporar reglas lógicas en el modelo, se facilita la comprensión de las decisiones del modelo, ya que las explicaciones pueden basarse en las reglas lógicas utilizadas. 3. **Métodos y Enfoques:** El artículo detalla métodos específicos para integrar estas restricciones lógicas en el entrenamiento de modelos de aprendizaje profundo, destacando cómo estas técnicas se pueden aplicar en diferentes contextos y problemas. 4. **Resultados y Aplicaciones:** Los resultados experimentales muestran que los modelos con restricciones lógicas no sólo son más precisos en sus predicciones, sino que también proporcionan explicaciones más coherentes y comprensibles sobre su funcionamiento.
13	Explicabilidad Robusta: Un Tutorial sobre Métodos de Atribución Basados en Gradientes para Redes Neuronales Profundas.	Nielsen, I. E., Dera, D., Rasool, G., Bouaynaya, N., & Ramachandran, R. P. (2022). Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks. *IEEE Signal Processing Magazine*, *39*(4), 73–84. https://doi.org/10.1109/MSP.2022.3142719	**Nielsen et al. (2022)** presentan un tutorial sobre **métodos de atribución basados en gradientes** para **redes neuronales profundas**, abordando la **explicabilidad robusta** de los modelos de aprendizaje profundo. **Aspectos Clave del Artículo:** 1. **Métodos de Interpretabilidad Basados en Gradientes:** ○ El artículo explora técnicas que utilizan **gradientes** para **asignar la responsabilidad** de las decisiones del modelo a las **características de entrada**. ○ Estas técnicas permiten entender cómo cada característica contribuye a la predicción realizada por el modelo. 2. **Enfoques Detallados:** ○ Se discuten varios métodos, incluyendo la **propagación de gradientes** y técnicas derivadas, que se utilizan para calcular la importancia de las características en función de los gradientes de la función de pérdida con respecto a las entradas. 3. **Beneficios y Aplicaciones:** ○ Los métodos basados en gradientes ofrecen **explicaciones intuitivas** sobre cómo las características afectan las decisiones del modelo, lo que facilita la interpretación y la **confianza** en el modelo. ○ Estos enfoques son particularmente útiles en **aplicaciones críticas** donde la comprensión de la influencia de las características es crucial para la toma de decisiones.

			4. **Tutorial y Aplicación Práctica:**El artículo proporciona un tutorial práctico, guiando a los lectores a través de la implementación e interpretación de los métodos basados en gradientes.Se incluyen ejemplos y casos de estudio que muestran cómo aplicar estos métodos en la práctica.
14	Explicando la caja negra: un enfoque contrafactual	Singla, S., Eslami, M., Pollack, B., Wallace, S., & Batmanghelich, K. (2023). Explaining the black-box smoothly—A counterfactual approach. *Medical Image Analysis, 84*, 102721. https://doi.org/10.1016/j.media.2022.102721	**Singla et al. (2023)** presentan el **"BlackBox Counterfactual Explainer"**, un modelo innovador para **explicar decisiones de clasificación en imágenes médicas**. Este enfoque contrafactual se compara con otros métodos tradicionales, como los mapas de saliencia. **Aspectos Clave del Artículo:** 1. **Problema con Enfoques Tradicionales:** ○ Los métodos tradicionales, como los mapas de saliencia, a menudo no explican de manera efectiva **cómo las características en regiones anatómicas específicas** influyen en la decisión del clasificador de imágenes médicas. 2. **Modelo Propuesto:** ○ El **"BlackBox Counterfactual Explainer"** se presenta como una solución para superar las limitaciones de los enfoques tradicionales, proporcionando explicaciones que revelan cómo diferentes características contrafácticas afectan la decisión del clasificador. 3. **Experimento Comparativo:** ○ Se realizó un experimento con residentes de radiología diagnóstica, comparando diferentes estilos de explicación: ■ Sin explicación ■ Mapas de saliencia ■ Explicación con **cycleGAN** ■ Explicación **contrafactual** ○ Se evaluaron varios aspectos, como la **comprensibilidad**, la **justificación de la decisión del clasificador**, la **calidad visual**, la **preservación de la identidad** y la **utilidad general** de cada tipo de explicación. 4. **Resultados y Conclusiones:** ○ El enfoque contrafactual demostró ser el único método que **mejoró significativamente la comprensión de los usuarios** sobre las decisiones del clasificador en comparación con la línea base sin explicación. ○ Este método ofrece una explicación más clara y útil para los usuarios, facilitando una mejor comprensión de cómo las decisiones se basan en características específicas de las imágenes.
15	Modelos de cuello de botella post-hoc	Yuksekgonul, M., Wang, M., & Zou, J.Y. (2022). Post-hoc Concept Bottleneck Models. *ArXiv, abs/2205.15480*.	**Yuksekgonul, Wang, y Zou (2022)** introducen los **Modelos de Cuello de Botella Conceptual (CBMs)** como una técnica para **mejorar la interpretabilidad de modelos de aprendizaje profundo** sin

			comprometer la precisión. **Aspectos Clave del Artículo:** 1. **Concepto de Cuello de Botella:** ○ Los CBMs se basan en la idea de **introducir una capa de cuello de botella conceptual** dentro de la arquitectura del modelo. Esta capa captura conceptos intermedios que se cree que son importantes para la toma de decisiones del modelo. 2. **Objetivo:** ○ La principal ventaja de los CBMs es que **revelan qué conceptos son relevantes para una predicción** específica, proporcionando una visión más clara de cómo el modelo llega a sus conclusiones. 3. **Beneficios:** ○ Los CBMs permiten una **interpretación más transparente** al identificar y destacar los conceptos clave que influyen en la predicción del modelo. ○ Importante destacar que esta mayor interpretabilidad **no sacrifica la precisión del modelo**, ya que los CBMs están diseñados para mantener la exactitud de las predicciones mientras se mejora la comprensión de los factores que influyen en ellas.
16	sMRI-PatchNet: Una Nueva Red Eficiente Explicable Basada en Parches para el Diagnóstico de la Enfermedad de Alzheimer con Resonancia Magnética.	Zhang, X., Han, L., Han, L., Chen, H., Dancey, D., & Zhang, D. (2023). sMRI-PatchNet: A Novel Efficient Explainable Patch-Based Deep Learning Network for Alzheimer's Disease Diagnosis With Structural MRI. *IEEE Access, 11,* 108603-108616. https://doi.org/10.1109/ACCESS.2023.3321220.	**Zhang et al. (2023)** presentan **sMRI-PatchNet**, una red neuronal profunda y explicativa diseñada para **el diagnóstico de la Enfermedad de Alzheimer (AD)** mediante resonancia magnética estructural (sMRI). Esta red se distingue por su enfoque en la eficiencia y la capacidad de explicación, y se compone de dos componentes clave: 1. **Selección de Parches Discriminativos:** ○ **sMRI-PatchNet** utiliza un **método rápido y eficiente para seleccionar parches de imagen** que son relevantes para la clasificación de AD. Esto optimiza el proceso de identificación de características patológicas en las imágenes. 2. **Extracción de Características y Clasificación:** ○ La red emplea una **estructura avanzada para extraer características profundas** de los parches seleccionados y clasificar la presencia de AD y la conversión de deterioro cognitivo leve (MCI). **Resultados Experimentales:** ● **Precisión y Rendimiento:** ○ Evaluado en **conjuntos de datos reales**, **sMRI-PatchNet** demostró una **identificación efectiva de ubicaciones patológicas** con una reducción significativa en el número de parches utilizados, lo que

			se traduce en **mejor rendimiento en términos de precisión**, **rendimiento computacional** y **generalización**. ● **Comparación con Métodos del Estado del Arte:** ○ El método superó a los enfoques existentes en términos de **precisión y eficiencia**, proporcionando un avance significativo en la detección y diagnóstico de la Enfermedad de Alzheimer.
17	Aprendizaje Contrastivo Supervisado para la Comparación Interpretativa de Documentos Largos	Jha, A., Rakesh, V., Chandrashekar, J., Samavedhi, A., & Reddy, C. K. (2023). Supervised contrastive learning for interpretable long-form document matching. *ACM Transactions on Knowledge Discovery from Data, 17*(2), Article 27. https://doi.org/10.1145/3542822	El artículo presenta un modelo denominado **CoLDE (Contrastive Long Document Encoder)**, diseñado para mejorar la correspondencia semántica en documentos largos, como artículos científicos, documentos legales y patentes. Los enfoques actuales tienden a enfocarse en documentos cortos y enfrentan dificultades con documentos extensos debido a: ● **Contextos variables**: Una misma palabra puede tener diferentes significados en diferentes partes del documento. ● **Secciones similares pero no idénticas**: Las secciones de texto pueden ser similares entre documentos, pero otras partes pueden diferir significativamente. ● **Medidas de similitud limitadas**: Las medidas de similitud actuales a menudo no capturan la heterogeneidad del contenido a lo largo del documento. CoLDE aborda estos desafíos utilizando una arquitectura basada en transformadores con incrustaciones posicionales únicas y una capa de atención por fragmentos, combinada con un aprendizaje contrastivo supervisado. Captura la similitud en tres niveles: 1. **Similitud general entre documentos**. 2. **Similitud entre secciones** dentro de un documento y entre documentos. 3. **Similitud entre fragmentos** dentro del mismo documento y con otros documentos. Estos niveles detallados de puntuación de similitud mejoran la interpretabilidad del modelo. CoLDE fue evaluado en tres conjuntos de datos de documentos largos: publicaciones del ACL Anthology, artículos de Wikipedia y patentes del USPTO. Superó a los métodos existentes en la tarea de correspondencia de documentos, mostrando robustez frente a variaciones en la longitud del documento y perturbaciones del texto. El código del modelo está disponible públicamente.
18	Un Marco de Aprendizaje Contrastivo Multi-expertos Centrado en Preguntas para Mejorar la Precisión y la Interpretabilidad de los Modelos de Rastreo de Conocimiento Secuencial Profundo	Zhang, H., Liu, Z., Shang, C., Li, D., & Jiang, Y. (2024). A question-centric multi-experts contrastive learning framework for improving the accuracy and interpretability of deep sequential knowledge tracing models. *ACM Transactions on Knowledge Discovery from Data*. https://doi.org/10.1145/3674840	El **rastreo del conocimiento (KT)** es crucial para predecir el rendimiento futuro de los estudiantes mediante el análisis de sus procesos de aprendizaje históricos. Aunque las redes neuronales profundas (DNN) han demostrado su potencial para abordar el problema de KT, enfrentan desafíos significativos: ● **Modelado de la información individual de las preguntas**: La adquisición de conocimiento por parte de los estudiantes puede variar considerablemente entre preguntas que involucran los mismos componentes de conocimiento (KCs). ● **Interpretación de los resultados de predicción**: Los modelos de KT basados en aprendizaje profundo necesitan resultados que los profesores puedan entender y aplicar en sus estrategias

			educativas. Para enfrentar estos desafíos, se propone un marco de aprendizaje contrastivo multi-expertos centrado en preguntas, denominado **Q-MCKT**. Este marco modela el estado de adquisición de conocimiento de los estudiantes a nivel de preguntas y conceptos mediante una técnica de mezcla de expertos, que captura un estado de conocimiento más robusto y preciso. Además, introduce una tarea de aprendizaje contrastivo centrada en preguntas para mejorar las representaciones de preguntas con menos interacciones y utiliza una capa de predicción basada en la teoría de respuesta al ítem para generar resultados más interpretables.
19	Identificación de las Necesidades de Explicación de los Usuarios Finales: Aplicación y Extensión del Banco de Preguntas de XAI	Sipos, L., Schäfer, U., Glinka, K., & Müller-Birn, C. (2023). Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank. In Mensch und Computer 2023 (MuC '23), September 03–06, 2023, Rapperswil, Switzerland (pp. 1–6). ACM, New York, NY, USA. https://doi.org/10.1145/3603555.360851	El artículo propone un enfoque centrado en el usuario para mejorar la explicabilidad en inteligencia artificial, denominado **Human-Centered Explainable Artificial Intelligence (HC-XAI)**. Mientras que las explicaciones actuales de IA suelen centrarse en la transparencia algorítmica, a menudo no satisfacen las necesidades de usuarios sin experiencia en IA. El estudio utiliza y extiende el **XAI Question Bank (XAIQB)**, una herramienta diseñada para identificar las necesidades de explicación de los usuarios. Se identificaron deficiencias en el XAIQB y dificultades en su aplicación práctica. En respuesta, los autores añadieron 11 preguntas nuevas y mejoraron las descripciones de las existentes, con el fin de facilitar su uso en el ámbito de la Interacción Humano-Computadora (HCI).
20	Abordando Métodos de Inteligencia Artificial Explicable en el Diagnóstico de la Anemia por Deficiencia de Hierro Usando Parámetros Sanguíneos	Ponnusamy, U., D. D. B. S., & Sampathila, N. (2023). Approaching explainable artificial intelligence methods in the diagnosis of iron deficiency anemia using blood parameters. In *2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)* (pp. 201-206). Manipal, India. https://doi.org/10.1109/ICRAIS59684.2023.10367126	La anemia por deficiencia de hierro es un trastorno de salud que se diagnostica mediante la observación de parámetros sanguíneos. Este proceso puede ser tedioso y propenso a errores cuando se realiza manualmente por profesionales de la salud. El artículo propone un método para mejorar la comprensión del impacto de los parámetros sanguíneos en el diagnóstico de la anemia. Se emplean métodos de *machine learning* para clasificar los datos y herramientas de *inteligencia artificial explicable (XAI)* para evaluar el impacto de los atributos, proporcionando así transparencia y confianza en los modelos utilizados. XAI asegura equidad, responsabilidad y claridad en las decisiones del modelo. Los modelos presentaron una alta precisión, alcanzando entre el 80% y el 100%.
21	IA Explicable para la Medicina por el intérprete de código de ChatGPT.	Kenta, K., Kitamura, M., Irvan, M., & Shigetomi Yamaguchi, R. (2023). XAI for Medicine by ChatGPT Code interpreter. In *2023 5th International Conference on Big-data Service and Intelligent Computation (BDSIC 2023)*, October 20-22, 2023, Singapore, Singapore. ACM, New York, NY, USA, 7 pages.	El estudio presenta un método para implementar *IA explicable* (XAI) en tareas médicas utilizando el intérprete de código de ChatGPT. Se propone un nuevo método denominado **Code Base Prompt (CBP)**, que emplea la función de ejecución de código en Python de ChatGPT para hacer más transparentes los procesos de toma de decisiones en textos médicos. Además, se introduce un sistema de evaluación de la explicabilidad llamado **Medical Algorithm Presentation Criteria (MAPC)**, que evalúa la alineación con el proceso de comprensión humana mediante cinco factores. En una comparación entre CBP y un enfoque basado en texto denominado **Text Base Prompt (TBP)**, se aplicó un algoritmo de clasificación de insuficiencia cardíaca a textos de reportes de casos en tres

		https://doi.org/10.1145/3633624.3633629	artículos médicos. Los resultados mostraron que el enfoque CBP ejecutó correctamente el código Python y cumplió con los cinco factores de MAPC en todos los casos. En contraste, el enfoque TBP no logró ejecutar el código ni cumplir con la mayoría de los factores.
22	Interpretación de Modelos de Aprendizaje Automático de Caja Negra para Conjuntos de Datos de Alta Dimensión	Karim, M. R., et al. (2023). Interpreting black-box machine learning models for high dimensional datasets. *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, 1-10. https://doi.org/10.1109/DSAA60987.2023.10302562	El artículo presenta un enfoque en varias etapas para mejorar la interpretabilidad de los modelos de caja negra en el contexto de conjuntos de datos de alta dimensionalidad. El proceso y la justificación de cada etapa son los siguientes: 1. **Entrenamiento del Modelo de Caja Negra**: Se entrena inicialmente un modelo de caja negra (como una red neuronal profunda) utilizando el conjunto de datos completo con todas sus características. Este modelo tiene la capacidad de aprender representaciones complejas y no lineales de los datos, que modelos más simples podrían no captar. 2. **Descomposición e Identificación de Características Importantes**: Después de entrenar el modelo de caja negra, se aplican técnicas de sondeo y perturbación para descomponer el modelo e identificar las características más influyentes (denominadas top-k características). El sondeo implica modificar sistemáticamente las características para observar cómo cambian las predicciones del modelo, mientras que la perturbación añade ruido a ciertas características para evaluar su impacto en el rendimiento del modelo. 3. **Entrenamiento de un Modelo Interpretable**: Con las características más importantes identificadas, se entrena un modelo interpretable, como un árbol de decisión o una regresión logística, utilizando solo ese subconjunto reducido de características. Este modelo interpretable es más transparente y fácil de entender, lo que facilita la comprensión de cómo las características seleccionadas influyen en las decisiones del modelo de caja negra. 4. **Generación de Reglas de Decisión y Contrafactuales**: Finalmente, a partir del modelo interpretable, se derivan reglas de decisión y contrafactuales. Las reglas de decisión proporcionan una explicación clara de cómo el modelo llega a sus conclusiones basadas en las características más importantes. Los contrafactuales muestran cómo cambiar ciertas características podría alterar la predicción del modelo, ayudando a comprender mejor la toma de decisiones en situaciones específicas.
23	xAI: Un modelo de inteligencia artificial explicable para el diagnóstico de la EPOC a partir de imágenes de radiografía de tórax (CXR)	Ikechukwu, A. V., & Murali, S. (2023). xAI: An explainable AI model for the diagnosis of COPD from CXR images. In *2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS)* (pp. 1-6). Mangaluru, India.	La Enfermedad Pulmonar Obstructiva Crónica (EPOC) es una preocupación de salud mundial y la tercera causa principal de muerte. Aunque las pruebas de espirometría son el método definitivo para diagnosticar EPOC, su accesibilidad es limitada en regiones con pocos recursos. En contraste, las radiografías de tórax (CXR) son universalmente disponibles, lo que sugiere su potencial como herramientas de detección preliminar para la EPOC. Este estudio aplica algoritmos de aprendizaje profundo al conjunto de datos a gran escala VinDR-CXR para identificar la EPOC en sus etapas

		https://doi.org/10.1109/ICDDS59137.2023.10434619	iniciales utilizando CXR. Se emplea el conjunto de datos ChestX-ray14 para el preentrenamiento del modelo y luego se utiliza VinDR-CXR para el desarrollo y validación del modelo. Mediante aprendizaje por transferencia, se observó que el modelo Xception, con un recall del 98.2% obtenido mediante ajuste fino, fue superior al modelo ResNet50. Los mapas de calor de Grad-CAM (Mapeo de Activación de Clase por Gradiente) y SHAP (Explicaciones Aditivas de Shapley) proporcionan explicabilidad para los casos verdaderos positivos en el modelo Xception a partir de ambos conjuntos de datos.
24	Inteligencia Artificial Explicable e Imagenología Cardíaca: Hacia Modelos Más Interpretables	Salih, A., Boscolo Galazzo, I., Gkontra, P., Lee, A. M., Lekadir, K., Raisi-Estabragh, Z., & Petersen, S. E. (2023). Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models. *Circulation: Cardiovascular Imaging, 16*. https://doi.org/10.1161/CIRCIMAGING.122.014519	Las aplicaciones de inteligencia artificial han tenido éxito en diversos campos médicos, incluyendo la imagenología cardíaca. Sin embargo, muchos modelos de aprendizaje automático, especialmente los de aprendizaje profundo, se consideran "caja negra" debido a su falta de explicabilidad en los resultados. Esta falta de transparencia requiere la adopción de métodos de inteligencia artificial explicable (XAI) para hacer los resultados del modelo comprensibles para los usuarios finales. Aunque existen pocos estudios en imagenología cardíaca que emplean metodologías XAI, este artículo ofrece una revisión exhaustiva de las investigaciones actuales que utilizan estas técnicas. Además, proporciona directrices simples y completas sobre XAI, discutiendo problemas abiertos y posibles direcciones futuras en el ámbito de la imagenología cardíaca.
25	Una gama de enfoques de aprendizaje automático explicables e interpretables para estudios genómicos.	Conard, A. M., DenAdel, A., & Crawford, L. (2023). A spectrum of explainable and interpretable machine learning approaches for genomic studies. *Wiley Interdisciplinary Reviews: Computational Statistics, 15*, e1617. https://doi.org/10.1002/wics.1617	Este artículo revisa el espectro de transparencia en modelos, desde "cajas negras" hasta modelos interpretables, destacando su aplicación en estudios genómicos. Además, aborda la integración del conocimiento biológico para mejorar la interpretabilidad y discute los desarrollos futuros en este campo.
26	Una revisión sobre la inteligencia artificial explicable en medicina (XAI): Progreso reciente, enfoque de explicabilidad, interacción humana y sistema de puntuación	Sheu, R.-K., & Pardeshi, M. S. (2022). *A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system*. *Sensors, 22*(22), 8068. https://doi.org/10.3390/s22208068	Este artículo presenta un estudio detallado sobre la inteligencia artificial explicable (XAI) en medicina, abarcando mejoras en modelos, métodos de evaluación, conjuntos de datos abiertos y futuras innovaciones. Se destacan métodos recientes de XAI, tanto locales como globales, aplicados al preprocesamiento, algoritmos basados en conocimiento y aprendizaje automático interpretable. Además, se propone un enfoque de "usuario en el bucle" para mejorar la colaboración entre humanos y máquinas, generando soluciones más explicables. También se abordan las limitaciones actuales de las puntuaciones y calificaciones en XAI, proponiendo un nuevo sistema de recomendaciones y puntuación para mejorar la explicabilidad en el ámbito médico.
27	Grafos de Conocimiento como Herramientas para el Aprendizaje Automático	Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence, 302*, 103627. https://doi.org/10.1016/j.artint.	El artículo revisa el uso de grafos de conocimiento para mejorar la explicabilidad en el aprendizaje automático. Examina cómo integrar estos grafos para hacer que los modelos de IA sean más interpretables. Utilizando un marco analítico basado en una revisión sistemática, el estudio analiza cómo estos enfoques híbridos pueden mejorar la comprensión, precisión y reactividad de los sistemas, y discute sus

	Explicable: Una Revisión	2021.103627	limitaciones, como la gestión del ruido y la eficiencia en la extracción de conocimiento.
28	Explicaciones Locales Basadas en Conceptos con Retroalimentación (CLEF)	EL Shawi, R., & Al-Mallah, M. H. (2022). Interpretable local concept-based explanation with human feedback to predict all-cause mortality. *Journal of Artificial Intelligence Research, 75*. https://doi.org/10.1613/jair.1.14019	Se presenta un marco de explicabilidad local basado en conceptos con retroalimentación humana (CLEF) para predecir la mortalidad por cualquier causa. El CLEF es un enfoque novedoso y agnóstico al modelo que utiliza conceptos etiquetados por clínicos en lugar de características crudas. Mapea las características de entrada a conceptos intuitivos de alto nivel y descompone la evidencia de la predicción en estos conceptos. Además, genera explicaciones contrafactuales que sugieren los cambios mínimos en la explicación basada en conceptos que llevarían a una predicción diferente. El estudio muestra que la retroalimentación directa de los usuarios es más efectiva que otras técnicas para alinear los conceptos aprendidos con las definiciones de conceptos de la realidad.
29	Sobre la explicabilidad de los modelos profundos de procesamiento del lenguaje natural	El Zini, J., & Awad, M. (2023). On the explainability of natural language processing deep models. ACM Computing Surveys, 55(5), Article 103, 1–31. https://doi.org/10.1145/352975	El artículo revisa métodos de explicabilidad para modelos de Procesamiento de Lenguaje Natural (NLP), abordando la dificultad de interpretar redes profundas utilizadas como cajas negras. Presenta métodos que explican embeddings de palabras, el funcionamiento interno de los modelos y las decisiones finales. También evalúa enfoques actuales y propone direcciones futuras para mejorar la explicabilidad en NLP.
30	Una Revisión Exhaustiva y Aplicación de Modelos de Aprendizaje Profundo Interpretables para la Predicción de Reacciones Adversas a los Medicamentos	Dubey, S. A., & Pandit, A. A. (2022). A comprehensive review and application of interpretable deep learning model for ADR prediction. *International Journal of Advanced Computer Science and Applications (IJACSA), 13*(9). http://dx.doi.org/10.14569/IJACSA.2022.0130924	La seguridad de los medicamentos es crucial en la atención médica. Este estudio diseña un marco para revisar investigaciones sobre la detección y predicción de Reacciones Adversas a los Medicamentos (ADRs). Se recopilaron 172 artículos, que se clasificaron en temas de detección y predicción de ADRs. Se analizaron las fuentes de datos, algoritmos y métricas de evaluación, y se diseñó un marco de aprendizaje profundo que aborda brechas en modelos existentes. El modelo con dos capas ocultas mostró un rendimiento óptimo para la predicción de ADRs. Además, se usó un modelo sustituto global para mejorar la interpretabilidad. La arquitectura propuesta supera limitaciones previas y subraya la importancia de la detección temprana de ADRs en la industria de la salud.

4.1.2 Análisis Comparativo de Métodos de Explicabilidad en Inteligencia Artificial.

En el campo de la inteligencia artificial (IA), la comprensión de cómo y por qué un modelo llega a una decisión es esencial para fomentar la confianza en sus resultados y tomar decisiones informadas. La siguiente tabla (véase tabla 3), presenta una comparación de diversos métodos de explicabilidad en IA, destacando sus propuestas, ámbitos de uso, beneficios principales y características específicas.

Tabla 3. Comparación de Métodos de Explicabilidad en Inteligencia Artificial: Análisis de Modelos.

#	Método /Enfoque usado	Propuesta	Ámbito de Uso	Beneficios Principales	Problema abordado	Generalizable	Transformación de Datos	Ej	Comparación con enfoques anteriores
1	Equidad Contrafactual Contrastiva	Introduce un enfoque de equidad contrastiva contrafactual para mejorar la justicia en decisiones algorítmicas.	Concesión de préstamos. Selección de candidatos. Evaluación crediticia.	- Mitigación de sesgos en las decisiones algorítmicas. - Aumento de la transparencia y explicabilidad de los modelos de IA. - Mejora de la confianza de los usuarios en los sistemas automatizados. - Capacidad de aplicar el enfoque a cualquier modelo de IA debido a su naturaleza agnóstica del modelo.	Explicaciones contrafactuales	SI	NO	NO	Sin mitigación y aumento de datos contrafácticos.
2	Aprendizaje Federado (Federated Learning, FL)	Caracterizar y explicar anomalías detectadas por modelos de detección de intrusiones basados en ML no supervisado, aplicados a entornos distribuidos como IoT.	Entornos distribuidos	Mejora la interpretación y eficacia de los sistemas de detección de intrusiones y seguridad en redes, aprovechando la distribución de datos.	Explicación de resultados	SI	SI	SI	SHAP (SHapley Additive exPlanation). LEMNA
3	RoCourseNet (Red de Optimización Tri-nivel, Entrenamiento Adversarial, Algoritmo de Cambio Virtual de Datos (VDS))	Clasificación estructurada de los elementos explicativos dentro del marco de RoCourseNet,utilizando técnicas como Virtual Data Shift (VDS) y CounterNet.	Mejora de la robustez de las explicaciones contrafactuales en modelos predictivos frente a desplazamientos de datos.	Generación de explicaciones contrafactuales válidas y robustas, mejorando la transparencia y generalización de modelos 'caja negra'.	Explicaciones contrafactuales.	SI	SI	SI	CounterNet, ROAR, y RB.
4	Red Neuronal Convolucional Multi-categoría Consciente del Espacio (SAMCNet)	- Caracterizar información espacial. -Descomponer características espaciales antes de aplicar la operación EdgeConv. - Subred de priorización de pares de puntos -Agregar información a través de todos los puntos vecinos mediante una función asimétrica.	Oncología, Farmacología, Biomédica, Paleontología, Ecología, Epidemiología.	- Mayor precisión en la predicción. - Eficiencia computacional. - Capacidad de descubrimiento de patrones.	Explicabilidad del Modelo	SI	SI	SI	PointNet; DGCNN (Dynamic Graph Convolutional Neural Network);SRNet (Spatial Relation Network).
	Explicaciones	Mejorar la interacción	Sistemas de	-Mejora del rendimiento	Explicabilidad	SI	NO	SI	*aIDS*, Action

5	basadas en subobjetivos (Subgoal-Based Explanations)	del usuario novato con sistemas IDS (Intelligent Decision Support) que pueden recomendar acciones subóptimas o fallar.	Soporte de Decisión Inteligente.	del usuario. -Distinción entre recomendaciones óptimas y subóptimas -Rendimiento robusto en caso de fallos de IDS	del Resultado				Recommendations from IDS. EC L C, Causal-Link-Chain Explanations.,
6	LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations)	Enfoque de interpretación de modelos CNN (red neuronal convolucional) aplicados a la identificación de defectos de software	Predicción de defectos de software.	Mejora la eficiencia y precisión en la detección de defectos de software, ofreciendo explicaciones tanto locales como globales.	Explicabilidad del modelo	SI	NO	SI	Deep Representation and Ensemble Learning. DAECNN-JDP. Transfer CNN Model.
7	Uso de Partes Prototípicas (PPs) y ProtoPNet para mejorar la interpretabilidad del modelo mediante visualizaciones claras.	Clasificación estructurada y jerárquica de los elementos explicativos del modelo (PPs).	Diagnóstico asistido por computadora (CADx) para reconocimiento de piedras renales in-vivo.	Mejora en la precisión y explicabilidad en la clasificación de cálculos renales, incrementando la confianza de los especialistas médicos.	Explicabilidad del modelo	SI	NO	SI	ProtoPNet se comparó con las CNN base en términos de precisión, precisión media, puntuación F1 y otras métricas de rendimiento bajo condiciones IID y OOD.
8	Regla difusa (Fuzzy Rule)	Sistemas explicativos basados en reglas difusas para DNNs	Medicina, finanzas y otras ciencias aplicadas.	Mejoras significativas en la interpretabilidad de modelos complejos, alta fidelidad en la replicación del comportamiento de DNNs, y eficiencia mediante el uso de un número reducido de reglas y características.	Explicabilidad del modelo	SI	SI	SI	Chi_FRBCS y ECLAIRE
9	Metodología de explicabilidad XAI organizada en cuatro ejes principales.	Proporcionar una visión general del estado actual y las tendencias en XAI, dividiendo las técnicas en cuatro ejes: (i) Explicabilidad de datos, (ii) Explicabilidad del modelo, (iii) Explicabilidad post-hoc (iv) Evaluación de explicaciones.	Data scientists y desarrolladores Expertos en dominio y consumidores finales Aplicaciones generales	-Mejora de la transparencia y comprensión. -Alineación con las necesidades del usuario final. -Diseño de sistemas explicables. -Reducción de costos y esfuerzos. -Mejora de la confianza y aceptación.	Explicación del modelo	SI	NO	SI	
10	Explicaciones Lógicas Basadas en Entropía de Redes Neuronales. (Entropy-Based Logic Explanations of	Criterio basado en entropía (Entropy-based criterion)	Aplicación de lógica de primer orden (First-Order Logic).	Proporciona explicaciones formales y concisas.	Explicabilidad del modelo	SI	NO	SI	Redes ψ, Árboles de decisión, Listas de Reglas Bayesianas.

	Neural Networks).								
11	Redes Explicadas por Lógica (LENs)	Logic Explained Networks (LENs) se basan en el uso de lógica de primer orden.	Médicos, financistas, usuarios no expertos, científicos y académicos.	Interpretabilidad Mejorada, Flexibilidad en la Granularidad de las Explicaciones.	Explicabilidad del Modelo, Extracción de Reglas, Explicaciones Locales.	SI	SI	SI	LIME (Explicaciones Locales Interpretables Modelo-Agnósticas), SHAP (Explicaciones Aditivas de Shapley), Maximización de Activación, Mapas de Saliencia, SP-LIME, Árboles de Decisión, Listas de Reglas Bayesianas (BRL) DeepRED.
12	Restricciones lógicas.	Integración de restricciones lógicas en la función de pérdida o en la salida del modelo. para mejorar la interpretación..	Aplicaciones diversas .	Garantiza el ajuste a restricciones específicas, mejora la interpretación del modelo y su robustez frente a casos adversarios.	Diseño Transparente	SI	NO	SI	KBANN (Knowledge-Based Artificial Neural Networks). CILP (Computational Inductive Logic Programming). LTN (Logic Tensor Networks). NeurASP (Neural Answer Set Programming). Iterative Rule Distillation, etc
13	Atribución Basada en Gradiente.	Proporcionar una visión general de los métodos de interpretabilidad basados en gradientes, y evaluarlos en términos de robustez.	Diagnóstico médico asistido por IA. Sistemas de conducción autónoma. Aplicaciones críticas. Investigación y desarrollo.	Rapidez.Robustez.Evaluación de la Carga de Decisión.	Explicabilidad del modelo.	SI	NO	SI	Gradiente de entrada (Gradient Input). IG (Integrated Gradients). Mapas de saliencia. DeconvNet. Retropropagación del gradiente. Guided Backpropagation con Grad-CAM. DeepLIFT. SmoothGrad.
14	Explicador Contrafactual para Cajas Negras	Uso de Redes Generativas Adversariales (GANs). Preservación de Detalles Esenciales. Métricas Cuantitativas Clínicamente Relevantes.	Diagnóstico por imágenes médicas.	Mejora de la Transparencia. Preservación de Detalles Esenciales. Métricas Clínicamente Relevantes. Identificación de	Explicación contrafactual	SI	SI	SI	Mapas de saliencia y cycleGAN

		Evaluación. Benchmarking y Transparencia.		Características Relevantes. Mejora de la Toma de Decisiones Médicas. Establecimiento de un Benchmark.					
15	Cuello de Botella Conceptual Post-hoc (P-CBM) y Cuello de Botella Híbrido (HP-CBM)	P-CBM (Post-hoc Concept Bottleneck Model): Convertir cualquier modelo preentrenado en un modelo de cuello de botella conceptual. HP-CBM (Hybrid Post-hoc Concept Bottleneck Model): Combinar un modelo de cuello de botella conceptual con un modelo residual para mejorar la precisión sin perder interpretabilidad	Diagnóstico médico.	Versatilidad sin Pérdida de Desempeño. Eliminación de las Etiquetas de Conceptos durante el Entrenamiento. Incremento en el Rendimiento sin Ajustes Adicionales.	Explicaciones Locales. Explicación de Reglas.	SI	NO	SI	Concept Bottleneck Models (CBMs) tradicionales.
16	sMRI-PatchNet	Selección Automática de parches discriminativos. Reducción de la complejidad computacional. Automatización en la extracción de regiones de interés (ROI) Interpretación y explicabilidad. Comparación con Métodos Tradicionales	Diagnóstico médico.	Mejora en la precisión del diagnóstico. Reducción significativa en el número de parches utilizados. Mejora en el rendimiento computacional. Mayor generalizabilidad. Interpretabilidad mejorada.	Explicación del modelo	SI	NO	SI	SVM, LDA y KNN
17	CoLDE (Contrastive Long Document Encoder) (Codificador Contrastivo de Documentos Largos)	Aprendizaje Contrastivo Supervisado. Atención Chunkwise Multi-cabezal. Estructura Jerárquica para Documentos Largos.	Investigación académica, análisis de patentes, búsqueda automática de documentos, gestión de enciclopedias digitales, y revisión de informes técnicos y científicos.	Interpretación Detallada. Mejora en la Explicabilidad. Muy buen rendimiento. Flexibilidad y Escalabilidad.	Explicación del modelo	SI	SI	SI	DSSM, ARC-I, Hierarchical Attention Networks (HAN), Siamese-BERT (S-BERT), SMITH y S-LONG.
18	Q-MCKT (Question-centric Multi-experts Contrastive Learning framework for	Modelo centrado en Preguntas. Uso de Múltiples Expertos. Aprendizaje Contrastivo Centrado en Preguntas.	Rastreabilidad del conocimiento en el ámbito educativo.	Mejora de la Precisión en la Predicción: - Optimización del Estado de Adquisición de Conocimiento. -Aprendizaje Contrastivo Fino.	Explicación del modelo	NO	SI	SI	LIME, SHAP, Grad-CAM, Integrated Gradients, Anchors y Counterfactual Explanations.

	Knowledge Tracing).	Capa de Predicción Basada en la Teoría de Respuesta a Ítems (IRT).		Reducción del Sobreajuste: Interpretabilidad de los Resultados: -Predicciones Basadas en IRT (teoría de respuesta al ítem). Robustez y Adaptabilidad. Mejora en el Manejo de Datos Desbalanceados.					
19	Análisis cualitativo a través de la técnica *think-aloud*, combinada con la extensión y mejora de XAIQB para abordar las necesidades de explicación de los usuarios finales.	Mejorar y extender el XAIQB para abordar mejor las necesidades de explicación de los usuarios finales en contextos específicos de uso, mediante la incorporación de nuevas preguntas y descripciones.	Usuarios finales	Personalización de Explicaciones. Mejora de la Transparencia. Facilita la interacción con el Sistema. Soporte para Toma de Decisiones Informadas.	Explicación del modelo	SI	NO	SI	Extensión de XAIQB con 11 nuevas preguntas y descripciones ampliadas.
20	Abordando Métodos de Inteligencia Artificial Explicable en el Diagnóstico de la Anemia por Deficiencia de Hierro Usando Parámetros Sanguíneos	Clasificación de Datos mediante Machine Learning. Explicabilidad a través de Herramientas XAI Utilizar el gráfico de enjambre de abejas para visualizar y explicar cómo cada atributo del hemograma completo contribuye al diagnóstico de la anemia	Diagnóstico Médico	Diagnóstico eficiente. Reducción de errores. Transparencia y confianza. Alta precisión. Explicabilidad de atributos.	Explicación del modelo	SI	NO	SI	Regresión Logística (Logistic Regression - LR) Random Forest (RF) Support Vector Machine (SVM) K-Nearest Neighbor (KNN)
21	Intérprete de código de ChatGPT con Código Base de Prompts (CBP)	Propuesta del Code Base Prompt (CBP) para hacer explicable el proceso de toma de decisiones de ChatGPT en textos médicos utilizando la ejecución de código en Python.	Diagnóstico Médico	Mejorar la explicabilidad de las decisiones de IA en textos médicos, cumplimiento de los criterios de presentación de algoritmos médicos (MAPC).	Explicación del modelo	SI	SI	SI	Text Base Prompt (TBP)
22	Modelado Sustituto con Importancia de Características y Técnicas de Sondeo	Combinar la capacidad predictiva de un modelo de caja negra con la interpretabilidad de un modelo más simple. Utilizar técnicas para identificar características relevantes y luego entrenar un modelo interpretable para aproximar y explicar las decisiones del modelo de caja negra.	Visión por computadora. Procesamiento del Lenguaje Natural. Tareas de Modelado Predictivo en Datos Tabulares.	Mejor la interpretabilidad de los modelos de caja negra al proporcionar explicaciones claras y detalladas sobre el impacto de las características en las predicciones, manejar datos de alta dimensionalidad de manera eficiente	Interpretación del modelo	SI	SI	SI	TabNet, XGBoost y SHAP.

23	Grad-CAM (Gradient-weighted Class Activation Mapping). SHAP (SHapley Additive exPlanations).	Combinar técnicas avanzadas de deep learning con métodos de explicabilidad (Grad-CAM y SHAP) para desarrollar un modelo de AI explicable que pueda mejorar el diagnóstico temprano de la EPOC utilizando imágenes de radiografía de tórax,	Diagnóstico médico de EPOC	Detección Temprana y Precisa de EPOC. Mejor Accesibilidad en Regiones con Recursos Limitados. Explicabilidad y transparencia. Confianza y Aceptación por Parte de los Profesionales de la Salud.	Explicación del modelo	SI	SI	SI	Método basado en ResNet50. Método basado en Xception. Grad-CAM. SHAP.
24	Grad-CAM (Gradient-weighted Class Activation Mapping). SHAP (SHapley Additive exPlanations). LIME (Local Interpretable Model-agnostic Explanations). SmoothGrad	Incrementar el Uso de XAI en Imágenes Cardíacas. Desarrollar y Utilizar Herramientas XAI de Código Abierto. Evaluación de Métodos XA:Evaluación Basada en la Aplicación. Evaluación Basada en el Humano. Evaluación Basada en la Funcionalidad. Validación Interna y Externa Rigurosa.	Diagnóstico médico: Imagenología Cardíaca.	Facilita la Comprensión de Resultados Clínicos. Transparencia y Comprensibilidad.	Explicación del modelo	SI	SI	SI	Modelos Clásicos (ML): Más interpretables (e.g., regresión lineal, árboles de decisión) pero con menor rendimiento en tareas complejas. - Modelos DL: Mayor rendimiento pero menos interpretables;
25	-SHAP (SHapley Additive exPlanations) -LIME (Local Interpretable Model-agnostic Explanations) complementadas con **Técnicas de Integración de Conocimiento Biológic**o: -Redes Parcialmente Conectadas Basadas en Anotaciones Biológicas. -Funciones de Pérdida Basadas en Principios Biológicos	Mejorar la transparencia y comprensión de los modelos de aprendizaje automático en el contexto de aplicaciones biomédicas	Diagnóstico Médico: Aplicaciones biomedicina, especialmente genómicas.	Mejora de la Interpretabilidad Avances en Técnicas de Explicabilidad. Técnicas de Integración de Conocimiento Biológico. Aplicaciones en Genómica y Biomedicina	Explicación del modelo. Interpretación del modelo.	SI	SI	SI	**Modelos "Caja Negra" Tradicionale**s: Redes Neuronales Profundas (DNNs) Máquinas de Soporte Vectorial (SVM). **Técnicas Post-hoc Tradicionale**s: Análisis de Importancia de Características. Visualizaciones de Activación
26	LIME-.(agnóstico al modelo)-GRAD-CAM: para el análisis de imágenes.	Enfoque en el Usuario y Colaboración Humano-Máquina. Sistema de Retroalimentación y Evaluación. Sistemas de Recomendación y Puntuación XAI.	Diagnóstico Médico.	Mejora de la Transparencia y Comprensión. Cumplimiento Legal y Ético. Avance en la Investigación y Desarrollo. Enfoque en la Colaboración Humano-Máquina.	Explicación del modelo	SI	SI	SI	SHAP (SHapley Additive exPlanations)-LRP (Layer-wise Relevance Propagation)

	XAI Recommendation System (XAI-RS) (sistema de recomendación) XAI Scoring System (XAI-SS)(sistema de puntuación)	Importancia de Implementar Soluciones Explicables en Medicina.		Retroalimentación Constructiva.					
27	Grafos de Conocimiento: KBX-systems (Knowledge-Based Explainable Systems)	Integración de Datos Diversos. Explicabilidad Mejorada. Aplicación Versátil Interacción con Usuarios	Recomendación de ítems, reconocimiento de imágenes, minería de datos y sistemas conversacionales, etc.	Explicaciones Simbólicas. Reutilización de Conocimientos. Adaptación a Diferentes Aplicaciones. Conocimiento Contextual. Apoyo en el aprendizaje.	Explicabilidad del modelo	SI	SI	SI	-Métodos post-hoc: LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations). -Métodos basados en reglas. -Explicabilidad mediante descomposición.
28	Explicaciones Locales Basadas en Conceptos con Retroalimentación (CLEF)	Incorporación de Retroalimentación Humana. Explicaciones Contrafactuales.	Diagnóstico Clínico.	Explicaciones claras y basadas en conceptos relevantes. Interacción con expertos para definir conceptos. Proporciona contrafactuales para explorar cambios en las predicciones. Mejora la detección de sesgos. Asegura la fidelidad del modelo. Apoyo a la toma de decisiones más informada en el ámbito clínico	Explicación del Modelo. Explicaciones Locales. Explicación del resultado. Explicaciones contrafactuales.	NO	SI	SI	Baseline Interactivo (AL). Baseline No Interactivo (LR).
29	Propone herramientas como exBERT y benchmarks como ERASER .	Explicabilidad de modelos profundos de procesamiento del lenguaje natural(NLP).	Procesamiento de lenguaje natural	Mejora de la Interpretabilidad en Modelos de NLP. Clasificación de Métodos de Explicación.	Explicación del modelo	SI	SI	SI	LIME/SHARP
30	SP-LIME (Sub-modular Pick Local Interpretable Model-Agnostic Explanations)	Desarrollo de un Modelo de Aprendizaje Profundo Interpretable. Uso de Técnicas de Explicación Locales.	Entorno médico. Predicción de Reacciones Adversas a los Medicamentos (ADRs)	Reducción de Riesgos para los Pacientes. Optimización de la Gestión del Tratamiento.	Explicación del modelo	SI	SI	SI	Máquinas de Soporte Vectorial (SVM). Árboles de Decisión. K-Nearest Neighbors (KNN)

La Tabla 4 proporciona una explicación de las columnas que se encuentran en la Tabla 3. Cada característica listada en esta tabla ayuda a comprender mejor los enfoques y técnicas

utilizados en los artículos analizados, permitiendo identificar de manera clara aspectos tales como el problema abordado, las técnicas específicas empleadas, los enfoques explicativos, la necesidad de transformaciones de datos, la inclusión de ejemplos ilustrativos y la comparación con enfoques previos.

Tabla 4. Abreviaturas correspondientes a la Tabla 3.

Característica	Descripción
Paper	El número hace referencia al artículo que se está analizando y cuyo detalle se encuentra en la tabla 2.
Problema Abordado	Categorías de problemas que los métodos de explicabilidad abordan, tales como ❖ Explicaciones contrafactuales ❖ Explicación de resultados ❖ Explicabilidad del modelo ❖ Explicabilidad del resultado ❖ Extracción de reglas ❖ Explicaciones locales ❖ Diseño transparente ❖ Interpretación del modelo
Método de explicabilidad	Técnicas específicas empleadas para proporcionar explicaciones, tales como: - **Equidad Contrafactual Contrastiva**: Mejora de la justicia y mitigación de sesgos - **Explicaciones basadas en subobjetivos**: Explicaciones centradas en objetivos específicos. - **LIME**: Explicaciones locales interpretables. - **SHAP**: Explicaciones basadas en Shapley values. - **Uso de Partes Prototípicas (PPs) y ProtoPNet**: Visualizaciones claras para mejorar la interpretabilidad. - **Regla difusa (Fuzzy Rule)**: Sistemas basados en reglas difusas para redes neuronales profundas. - **Metodología de explicabilidad XAI**: Organización de técnicas en cuatro ejes principales. - **Explicaciones Lógicas Basadas en Entropía**: Uso de lógica de primer orden para explicaciones. - **Redes Explicadas por Lógica (LENs)**: Modelos que explican sus decisiones a través de lógica. - **Restricciones lógicas**: Integración de restricciones lógicas en modelos de aprendizaje. - **Atribución Basada en Gradiente**: Técnicas que utilizan gradientes para explicar decisiones. - **Explicador Contrafactual para Cajas Negras**: Preserva detalles en explicaciones para modelos de caja negra.

	- **Cuello de Botella Conceptual Post-hoc (P-CBM)**: Optimización post-hoc de precisión y explicabilidad. - **Cuello de Botella Híbrido (HP-CBM)**: Uso combinado de técnicas para mejorar la explicación. - **Modelado Sustituto con Importancia de Características**: Modelos interpretables para explicar modelos complejos. - **Grad-CAM**: Visualización de activaciones de clases en redes convolucionales. - **SmoothGrad**: Mejora de visualizaciones de gradientes para mayor claridad. - **Explicaciones Locales Basadas en Conceptos con Retroalimentación (CLEF)**: Explicaciones basadas en conceptos con retroalimentación. - **SP-LIME**: Método modular de LIME con selección sub-molecular.
Enfoques de explicabilidad	Métodos y técnicas que se utilizan para abordar la explicabilidad en distintos contextos, tales como: - **Aprendizaje Federado (Federated Learning, FL)**: Técnicas que permiten entrenamiento de modelos distribuidos. - **RoCourseNet**: Red de optimización tri-nivel para entrenamiento adversarial y cambio virtual de datos. - **SAMCNet**: Red neuronal convolucional multi-categoría consciente del espacio. - **sMRI-PatchNet**: Red para análisis de imágenes cerebrales basada en parches. - **CoLDE**: Codificador de documentos largos contrastivo. - **Q-MCKT**: Framework centrado en preguntas para el rastreo de conocimientos. - **XAIQB**: Herramienta para abordar necesidades de explicación de usuarios finales. - **Código de ChatGPT con Código Base de Prompts (CBP)**: Ejemplos de uso de modelos de lenguaje para explicabilidad. - **Técnicas de Integración de Conocimiento Biológico**: Métodos que incorporan conocimiento biológico para la explicabilidad. - **Redes Parcialmente Conectadas Basadas en Anotaciones Biológicas**: Redes neuronales con anotaciones biológicas para mejor interpretación. - **Funciones de Pérdida Basadas en Principios Biológicos**: Pérdidas adaptadas a principios biológicos. - **XAI Recommendation System (XAI-RS)**: Sistema de recomendación basado en XAI. - **XAI Scoring System (XAI-SS)**: Sistema de puntuación para evaluar la explicabilidad. - **Grafos de Conocimiento: KBX-systems**: Sistemas explicables basados en grafos de conocimiento.
Generalizable	Indica si el enfoque de explicabilidad puede ser aplicado a diferentes modelos de caja negra de manera general.
Transformación de datos	Indica si el método requiere perturbaciones o permutaciones aleatorias del conjunto de datos original para generar explicaciones.

Ejemplos	Indica si el artículo proporciona ejemplos ilustrativos de cómo se aplica la explicabilidad.
Enfoques anteriores	Describe con qué métodos o enfoques previos se comparó el modelo propuesto, destacando mejoras o diferencias con técnicas existentes.

A continuación, se presentan los principales métodos y enfoques recuperados durante la revisión, que se detallan en la Tabla 3 y 4. Estos enfoques abordan diversas técnicas de XAI, cada una con aplicaciones y características únicas:

4.1.2.1 Desglose Detallado de Métodos y Enfoques en las Tablas 3 y 4.

A. **Métodos Basados en Lógica:**

a. Equidad Contrafactual Contrastiva: Se enfoca en reducir sesgos en decisiones basadas en datos, proporcionando explicaciones que contrastan escenarios hipotéticos. Ejemplo: En un modelo de crédito, se puede mostrar cómo variaciones en los ingresos de un solicitante afectarían la decisión.

b. Explicaciones Lógicas Basadas en Entropía de Redes Neuronales: Utiliza lógica de primer orden para ofrecer explicaciones formales y concisas. Ejemplo: En el diagnóstico médico, se puede explicar cómo ciertas características en una imagen influyen en la clasificación.

c. Redes Explicadas por Lógica (LENs)**:** Proporciona explicaciones jerárquicas y basadas en reglas para mejorar la interpretabilidad de modelos complejos. Ejemplo: LENs descomponen la decisión de un modelo de clasificación en una serie de reglas lógicas.

d. Restricciones Lógicas: Integra reglas lógicas en modelos para mejorar la interpretación y robustez de las decisiones. Ejemplo: En la clasificación de texto, las restricciones lógicas obligan al modelo a considerar términos clave.

B. **Métodos Basados en Datos Espaciales y Distribuidos:**

a. Aprendizaje Federado (Federated Learning, FL)**:** Permite la colaboración en la construcción de modelos de machine learning sin necesidad de centralizar los datos, mejorando la privacidad y la seguridad.

b. SAMCNet: Analiza información espacial en diversos dominios, como la oncología, para proporcionar interpretaciones más detalladas de los modelos. Ejemplo: SAMCNet puede analizar imágenes multiplexadas para identificar y clasificar células tumorales.

C. **Técnicas de Explicación Contrafáctica:**

   a. RoCourseNet: Mejora las explicaciones contrafácticas en modelos predictivos mediante una red de optimización tri-nivel y entrenamiento adversarial.

   b. Explicador Contrafactual para Cajas Negras: Utiliza Generative Adversarial Networks (GANs) para generar explicaciones en el diagnóstico de imágenes médicas. Ejemplo: GANs generan imágenes alternativas que muestran cómo cambios en características afectan la predicción del modelo.

D. **Métodos de Interpretabilidad Basados en Modelos y Enfoques Alternativos:**

   a. LIME (Local Interpretable Model-agnostic Explanations) **y** SHAP (SHapley Additive exPlanations**):** Ofrecen explicaciones locales e interpretables para una amplia variedad de modelos de machine learning.

   b. Uso de Partes Prototípicas (PPs) y ProtoPNet: Mejora la interpretabilidad del modelo mediante visualizaciones claras, especialmente en aplicaciones médicas.

   c. Regla Difusa (Fuzzy Rule)**:** Potencia la interpretabilidad en modelos complejos mediante la utilización de reglas difusas.

   d. Atribución Basada en Gradiente: Evalúa la robustez de los métodos de interpretabilidad basados en gradientes.

   e. Modelado Sustituto con Importancia de Características y Técnicas de Sondeo: Emplea modelos sustitutos para entender mejor el comportamiento de modelos más complejos.

E. **Métodos de Explicación Basados en Visualización:**

   a. Grad-CAM (Gradient-weighted Class Activation Mapping): Proporciona visualizaciones que destacan las regiones relevantes en imágenes para la clasificación de modelos.

   b. SmoothGrad**:** Mejora las visualizaciones basadas en gradientes al reducir el ruido en las atribuciones.

F. **Técnicas de Explicabilidad en Sistemas de Recomendación y Evaluación:**

   a. XAI Recommendation System (XAI-RS)**:** Desarrolla sistemas de recomendación con explicaciones interpretables.

b. XAI Scoring System (XAI-SS)**:** Proporciona sistemas de puntuación que evalúan la calidad de las explicaciones generadas.

G. **Técnicas de Explicación con Retroalimentación:**

a. Explicaciones Locales Basadas en Conceptos con Retroalimentación (CLEF)**:** Utiliza retroalimentación para ajustar y mejorar las explicaciones locales.

H. **Metodologías y Herramientas Adicionales:**

a. Metodología de Explicabilidad XAI: Organiza técnicas en cuatro ejes principales para una aplicación amplia de la explicabilidad.

b. SP-LIME (Sub-modular Pick Local Interpretable Model-Agnostic Explanations)**:** Ofrece una variante de LIME que mejora la selección de ejemplos explicativos.

c. Análisis cualitativo a través de la técnica think-aloud: Utiliza la técnica think-aloud para obtener retroalimentación cualitativa sobre la comprensión de las explicaciones.

d. Abordando Métodos de Inteligencia Artificial Explicable en el Diagnóstico de la Anemia por Deficiencia de Hierro Usando Parámetros Sanguíneos: Enfoca en aplicaciones específicas en el diagnóstico médico.

La revisión revela una evolución en los métodos de explicabilidad, con enfoques cada vez más sofisticados que buscan abordar las limitaciones de los modelos 'caja negra'. Los métodos actuales abarcan una variedad de técnicas, desde enfoques basados en lógica hasta métodos contrafácticos y de visualización. La diversidad en los enfoques refleja la complejidad de la tarea y la necesidad de soluciones específicas para diferentes aplicaciones.

4.1.2.2 Análisis de Aplicaciones y Contextos

El siguiente análisis, basado en los datos presentados en la Tabla 3 y 4, examina cómo diversos métodos de explicabilidad se aplican en diversos contextos. Estos incluyen la mejora de diagnósticos médicos, la detección de anomalías en sistemas de seguridad, la promoción de la equidad en decisiones automatizadas, y la interpretación de grandes volúmenes de texto. Las técnicas se revisan y agrupan según su ámbito de uso:

A. Aplicaciones Médicas y Biológicas:

a. **Papers Relevantes**: SAMCNet, sMRI-PatchNet, Entropy-Based Logic Explanations, Deep Prototypical-Parts (PPs), ProtoPNet.

b. **Ámbito:** Los métodos de explicabilidad en el ámbito médico, como SAMCNet para imágenes multiplexadas de inmunofluorescencia y sMRI-PatchNet para el diagnóstico de enfermedades neurológicas, son esenciales para mejorar la precisión diagnóstica y la interpretación de modelos complejos. Estas técnicas permiten una visualización clara de patrones y características importantes, facilitando la confianza de los médicos en las decisiones automatizadas y asegurando la transparencia en el proceso de diagnóstico.

B. Seguridad y Detección de Anomalías:

a. **Papers Relevantes**: Federated Learning (FL), RoCourseNet, Gradient-weighted Class Activation Mapping (Grad-CAM).

b. **Ámbito**: En el contexto de seguridad y detección de anomalías, los métodos de explicabilidad ayudan a interpretar las decisiones de los modelos en sistemas distribuidos y redes de alta seguridad. Federated Learning y RoCourseNet son útiles para asegurar la integridad y la interpretabilidad en entornos colaborativos, mientras que Grad-CAM proporciona información visual sobre las áreas relevantes en datos sensibles, como imágenes de vigilancia o logs de eventos.

C. Equidad y Transparencia en la Toma de Decisiones:

a. **Papers Relevantes:** Contrastive Counterfactual Fairness, Subgoal-Based Explanations, LIME, SHAP.

b. **Ámbito:** La equidad en las decisiones algorítmicas es crucial en aplicaciones que afectan a las personas, como la concesión de créditos o la selección de candidatos. Métodos como la Equidad Contrafactual Contrastiva y las Explicaciones Basadas en Subobjetivos abordan sesgos y promueven la justicia en las decisiones automatizadas. LIME y SHAP permiten una interpretación comprensible de los modelos, facilitando una mayor transparencia en las recomendaciones y decisiones algorítmicas.

D. Interpretación de Modelos de Texto y Documentos:

a. **Papers Relevantes**: CoLDE, Decision Tree (DT), Decision Rules (DR), Partial Dependence Plot (PDP).

b. **Ámbito**: En el análisis de textos extensos y documentos largos, técnicas como CoLDE proporcionan una interpretación detallada basada en la estructura del

documento. Otros métodos como Decision Trees y Partial Dependence Plots ayudan a descomponer y entender las decisiones de modelos en el análisis de texto, facilitando la comprensión de patrones y relaciones en grandes volúmenes de datos textuales.

E. Visualización y Explicaciones Basadas en Reglas:

a. **Papers Relevantes**: Fuzzy Rule-Based Explainer Systems (FRBES), ProtoPNet, Attention Mechanisms.

b. **Ámbito**: La visualización y las explicaciones basadas en reglas son útiles para interpretar modelos complejos en diversas aplicaciones. Los sistemas basados en reglas, como FRBES, y las técnicas de atención proporcionan explicaciones claras y accesibles sobre cómo los modelos procesan la información y toman decisiones, mejorando la comprensión general de los resultados del modelo.

La variedad de métodos de XAI presenta un amplio espectro de herramientas para mejorar la interpretabilidad y la transparencia en modelos de machine learning. Cada técnica ofrece ventajas y limitaciones específicas, y la selección adecuada depende del contexto de aplicación y los requisitos específicos de interpretabilidad. La integración de métodos lógicos, espaciales, y de visualización, junto con enfoques centrados en la equidad y la privacidad, proporciona un marco que favorece el desarrollo de sistemas de inteligencia artificial más confiables y transparentes.

4.1.2.3 Evaluación de Criterios de Selección y Eficacia de Aplicaciones

El siguiente cuadro (véase Tabla 5) proporciona una visión estructurada de cómo cada artículo se clasifica según los criterios de XAI propuestos por Vilone, G. y Longo, L. (2021), y cómo estos métodos abordan la explicabilidad.

Este análisis incluye:

- Problemas específicos: Los tipos de problemas que cada método aborda, como clasificación o regresión.
- Técnicas utilizadas: Los enfoques y métodos específicos empleados para proporcionar explicaciones.
- Características de los modelos y datos: Las características de los modelos de machine learning y los datos con los que se trabaja.

Cada entrada en la tabla representa un método específico extraído de los artículos seleccionados para esta revisión sistemática. La tabla proporciona información sobre:

1. Paper: El número hace referencia al artículo que se está analizando y cuyo detalle se encuentra en la tabla 2.
2. Escenario XAI: Si el método es post-hoc (aplicado después del entrenamiento) o ante-hoc (integrado durante el entrenamiento), y si es agnóstico (general) o específico (para un tipo particular de modelo).
3. Alcance XAI: La amplitud de las explicaciones proporcionadas, ya sea global (para el modelo entero) o local (para instancias individuales o subconjuntos).
4. Tipo de Problema XAI: El tipo de tarea que el método aborda, como clasificación o regresión.
5. Dato de Entrada XAI: El tipo de datos utilizados para el análisis, como numéricos, categóricos, imágenes, texto o series temporales.
6. Formato de Salida XAI: Cómo se presenta la explicación, ya sea mediante reglas, información textual, visualizaciones, o combinaciones de estos formatos.

La Tabla 5 proporciona una visión integral y estructurada de los métodos de XAI evaluados en los artículos seleccionados. Al clasificar los métodos según los criterios mencionados, la tabla permite:

- Comparar Enfoques de XAI: Evaluar cómo diferentes métodos se aplican en la práctica, su alcance, y el formato de salida que utilizan.
- Identificar la Aplicabilidad: Entender cómo cada método aborda la explicabilidad en distintos tipos de modelos de caja negra y en diversos contextos de datos.
- Facilitar la Comprensión: Ofrecer una visión clara de las características y capacidades de cada método, ayudando a seleccionar el enfoque más adecuado según las necesidades específicas de interpretabilidad y transparencia en modelos de machine learning.

Esta clasificación facilita la identificación de tendencias actuales y direcciones futuras en el desarrollo de técnicas XAI, ayudando a entender cómo los métodos contribuyen a la explicabilidad en inteligencia artificial.

Tabla 5. Clasificación de Métodos de XAI en Artículos Seleccionados según Criterios de Explicabilidad propuestos por Vilone, G. y Longo, L. (2021).

Paper	Escenario XAI	Alcance XAI	Tipo de Problema XAI	Dato de Entrada XAI	Formato de salida XAI
[1]	Post-hoc: Modelo agnóstico.	Global	Clasificación	Numérico/ Categórico	Información textual y numérica.
[2]	Post-hoc: Agnóstico	Local y Global	Clasificación	Numérico/ Categórico	Reglas/Información numérica/Visual

[3]	Post-hoc: Modelo agnóstico	Global	Clasificación	Numérico/ Categórico /Texto	Información numérica/ Información textual.
[4]	Post-hoc: Modelo específico	Global	Clasificación	Imágenes	Información Visual
[5]	Post-hoc: Modelo específico	Local	Clasificación	Numérico/ Categórico	Información textual
[6]	Post-hoc: Modelo específico	Local	Clasificación	Numérico/ Categórico	Información numérico y textual
[7]	Post-hoc: Modelo específico	Local	Clasificación	Imágen	Información Visual/Información numérica.
[8]	Post-hoc: modelo específico	Local y Global	Clasificación.	Numérico/ Categórico	Reglas difusas.
[9]	Post-hoc: Agnóstico	Local	Clasificación y Regresión	Numérico, Categórico, Texto, Visual	Explicaciones de características, análisis detallado, visualizaciones.
[10]	Post-hoc: Agnóstico	Global	Clasificación	Numérico/categórico, imágen y texto.	Información Numérica, Información Visual
[11]	Post-hoc: Agnóstico	Global, Local	Clasificación	Numérico/ categórico	Reglas, Información Numérica.
[12]	Ante-hoc: Agnóstico	Global, Local.	Clasificación	Numérico/Categórico, Imágenes	Información Numérica, información visual.
[13]	Post-hoc: Agnóstico	Global, Local	Clasificación	Imagen	Información Visual.
[14]	Post-hoc: Agnóstico	Global	Clasificación	Imagen	Información Visual.
[15]	Post-hoc: Agnóstico	Global	Clasificación	Numérico/Categórico, Imagen, Texto, Serie Temporal.	Reglas, Información Numérica, Textual y Visual
[16]	Post-hoc: específico	Global	Clasificación	Imagen	Información Visual
[17]	Ante-hoc	Global, Local	Clasificación	Texto	Información textual.
[18]	Post-hoc: Agnóstico	Local	Clasificación	Numérico/Categórico	Información Numérica, Información Textual
[19]	Ante-hoc	Global y Local	Clasificación/ Regresión	Numérico/Categórico, Imagen, Texto, Serie Temporal	Combinación de los Formatos

[20]	Post-hoc: Agnóstico	Local	Clasificación	Numérico/Categórico	Información Numérica. Información Visual.
[21]	Ante-hoc/ Post-hoc: Específico	Global/ Local	Clasificación/Regresión	Numerical/Categorical,Imagen,Texto,Serie Temporal	Reglas,Información Numérica, Información Textual, Información Visual.
[22]	Post-hoc: Agnóstico	Global/ Local	Clasificación y Regresión	Numérico/Categórico	Información Numérica
[23]	Post-hoc: Agnóstico	Global/ Local	Clasificación y Regresión	Imágenes	Información Visual
[24]	Post-hoc: Agnóstico/ específico	Global /Local	Clasificación y Regresión	Numérico/Categórico Imágen	Información Numérica Información textual Información visual
[25]	Post-hoc: Específico. Ante-hoc	Global/ Local	Clasificación	Numérico/Categórico	Información Numérica
[26]	Post-hoc: Agnóstico	Global	Clasificación	Numérico/categórico/ Imágenes/texto/serie temporal	Reglas/ Información numérica/Información textual/Información Visual/
[27]	Post-hoc: Agnóstico	Global	Clasificación	Numérico/Categórico/ Texto	Reglas/Información Numérica/Información Textual/Información Visual.
[28]	Post-hoc: Agnóstico	Local	Clasificación	Numérico/Categórico	Información Textual Información Contrafactual Información Visual
[29]	Post-hoc: Agnóstico.	Global: ERASER/ Local: exBERT	Clasificación	Texto	exBERT:Información Visual/textual
[30]	Post-hoc: Agnóstico	Local	Clasificación	Numérico/Categórico	Información Textual

La elección del método de explicación depende de factores como el momento en que se ofrece la explicación, su alcance y el tipo de problema abordado. A continuación, se presenta un análisis detallado que examina cómo se distribuyen y aplican estos métodos:

A. Escenarios de Explicación:

a. **Post-hoc:** La mayoría de los métodos proporcionan explicaciones después del entrenamiento del modelo. Ejemplos incluyen *Contrastive Counterfactual Fairness* ([1]), *Federated Explainability* ([2]), y *Deep Prototypical-Parts* ([7]).

Estos métodos ofrecen explicaciones una vez que el modelo está entrenado, ayudando a entender decisiones ya tomadas.

b. **Ante-hoc:** Menos comunes, los métodos ante-hoc integran explicaciones durante el entrenamiento del modelo. Ejemplos incluyen *Deep Learning with Logical Constraints* ([12]) y *Supervised Contrastive Learning* ([17]). Estos enfoques buscan mejorar la transparencia desde el inicio del entrenamiento.

B. Alcance de las Explicaciones:

a. **Global:** Algunos métodos proporcionan una visión global del modelo, como *Logic Explained Networks* ([11]), *Entropy-Based Logic Explanations* ([10]), y *Knowledge Graphs as Tools for Explainable Machine Learning* ([27]). Estos métodos ofrecen una visión completa de cómo funciona el modelo en general.

b. **Local:** Otros se centran en explicaciones para instancias específicas, como *LCNN* ([6]) y *Federated Explainability* ([2]). Estos enfoques ofrecen detalles sobre decisiones individuales del modelo.

c. **Combinado:** Métodos como *Fuzzy Rule-Based Explainer Systems* ([8]) y *Explainable Artificial Intelligence (XAI): What We Know and What is Left to Attain* ([9]) abordan tanto explicaciones globales como locales, proporcionando una visión completa y detallada.

C. Tipo de Problema:

a. **Clasificación:** La mayoría de los métodos están orientados a problemas de clasificación, como *SAMCNet* ([4]), *sMRI-PatchNet* ([16]), y *Post-hoc Concept Bottleneck Models* ([15]). Estos enfoques son comunes en aplicaciones donde se clasifican datos en categorías.

b. **Regresión:** Algunos métodos también abordan problemas de regresión, como *XAI: What We Know and What is Left to Attain* ([9]) y *A Survey on Medical Explainable AI* ([26]). Estos enfoques son útiles para predecir valores continuos.

4.1.2.4 Distribución de Métodos de XAI Según Categorías

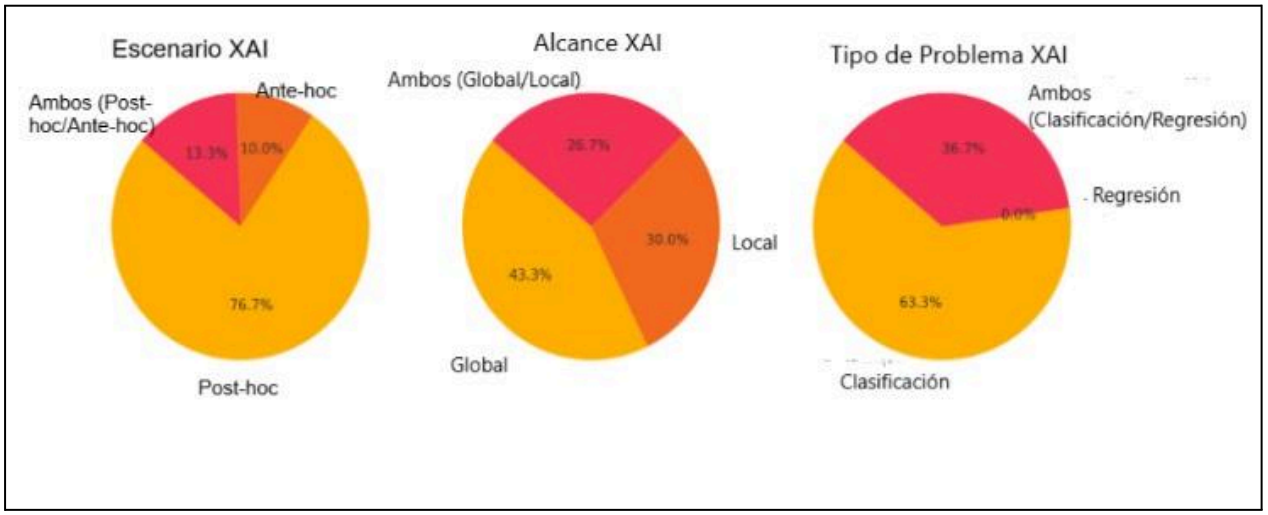
El Gráfico 1 ilustra la distribución de los enfoques de XAI en la literatura revisada, clasificados en tres categorías: Escenario XAI, Alcance XAI y Tipo de Problema XAI. Dando una perspectiva sobre las tendencias y enfoques actuales en el campo de XAI.

A. **Escenario XAI:** Esta categoría examina la proporción de métodos que utilizan enfoques Post-hoc, Ante-hoc, o una combinación de ambos. Según el Gráfico 1, una mayoría significativa del 76,7% de los métodos adoptan un enfoque Post-hoc (23 métodos), lo que indica una tendencia predominante hacia la explicación de modelos después de su entrenamiento. En comparación, el 13,3% emplea tanto enfoques Post-hoc como Ante-hoc (4 métodos), reflejando una flexibilidad en la aplicación de técnicas de explicabilidad. El 10% restante utiliza exclusivamente un enfoque Ante-hoc (3 métodos) , integrando la explicabilidad durante el proceso de entrenamiento del modelo.

B. **Alcance XAI:** Este aspecto aborda la extensión de las explicaciones proporcionadas, ya sea a nivel Global, Local, o ambos. El gráfico revela que el 43,3% de los métodos tienen un alcance global (13 métodos), ofreciendo explicaciones que abarcan el modelo completo. En contraste, el 30% se centra en explicaciones locales (9 métodos), que detallan instancias individuales o subconjuntos de datos. Un 26,7% de los métodos combinan ambos alcances, proporcionando una visión más completa al cubrir tanto explicaciones globales como locales (8 métodos).

C. **Tipo de Problema XAI:** Aquí se clasifica la proporción de métodos dirigidos a problemas de Clasificación, Regresión, o ambos. El Gráfico 1 muestra que el 63,3% de los métodos se especializan exclusivamente en problemas de Clasificación (19 métodos), mientras que el 36,7% abordan tanto Clasificación como Regresión (11 métodos), reflejando una menor concentración en problemas de Regresión.

Gráfico 1. Distribución en porcentajes de Métodos de XAI según Escenario, Alcance y Tipo de Problema.



D. Datos de Entrada y Formato de Salida:

a. **Datos de Entrada:** Los métodos manejan diversos tipos de datos, incluyendo numéricos, categóricos, imágenes y texto. Ejemplos son *Deep Prototypical-Parts* ([7]) para imágenes y *Contrastive Counterfactual Fairness* ([1]) para datos numéricos y categóricos.

b. **Formato de Salida:** Los formatos de salida varían, abarcando información textual, numérica, visualizaciones y reglas. Por ejemplo, *Fuzzy Rule-Based Explainer Systems* ([8]) proporcionan reglas difusas, mientras que *Deep Learning with Logical Constraints* ([12]) ofrece información numérica y visual.

E. Aplicaciones Específicas vs. Generales:

a. **Métodos Específicos:** Algunos enfoques están diseñados para aplicaciones concretas, como SAMCNet ([4]) para imágenes y sMRI-PatchNet ([16]) para el diagnóstico de Alzheimer."

b. **Métodos Generales:** Otros métodos son más agnósticos y aplicables a una variedad de contextos, como *Entropy-Based Logic Explanations* ([10]) y *A Comprehensive Review and Application of Interpretable Deep Learning Models for ADR Prediction* ([30]).

F. Tendencias y Direcciones Futuras:

a. **Enfoques Híbridos:** Existe una tendencia hacia métodos que combinan explicaciones post-hoc y ante-hoc para mejorar la transparencia y la comprensión del modelo. Ejemplos incluyen *Fuzzy Rule-Based Explainer Systems* ([8]) y *Logic Explained Networks* ([11]).

b. **Adaptación y Especificidad:** Los enfoques actuales buscan adaptarse a diferentes tipos de datos y problemas, reflejando un enfoque creciente en la personalización de las explicaciones para necesidades específicas.

El panorama de XAI está en expansión, con una rica variedad de métodos que ofrecen explicaciones para modelos de aprendizaje automático desde diferentes perspectivas y para diversos tipos de datos y problemas. La continua evolución y diversificación de estos métodos subraya la importancia de mejorar la transparencia y la interpretabilidad en la inteligencia artificial.

Capítulo 5

Conclusiones

5.1 Introducción

Este capítulo resume las conclusiones del estudio sobre Inteligencia Artificial Explicable (XAI), enfocado en la transparencia y la explicabilidad. La IA ha generado avances en la productividad, en la eficiencia e innovación (Baeza-Yates, 2024), pero también presenta un desafío crítico: la falta de transparencia en muchos sistemas. En contextos donde las decisiones dependen de algoritmos de IA, la explicabilidad es crítica para asegurar confianza, responsabilidad y aceptación pública.

En el ámbito de la IA, se distingue entre modelos interpretables, que ofrecen un proceso de decisión transparente, y modelos caja negra, que requieren explicaciones adicionales. Estas explicaciones deben cumplir objetivos cognitivo-conductuales, como promover la confianza, identificar sesgos y apoyar la toma de decisiones.

5.2 Principales Conclusiones

5.2.1 Necesidad de Explicabilidad en Modelos de IA

La explicabilidad es fundamental en aplicaciones de alto riesgo, como medicina, finanzas y seguridad, donde las decisiones automatizadas pueden tener un impacto significativo en la vida humana. Los modelos interpretables permiten una comprensión directa de cómo se toman las decisiones, mientras que los métodos diseñados para explicar modelos caja negra ayudan a construir confianza y garantizar la responsabilidad. La falta de explicabilidad en estos contextos puede llevar a una menor aceptación de las decisiones automatizadas y a una posible desconfianza en la tecnología.

5.2.2 Distinción entre Interpretabilidad y Explicabilidad

La interpretabilidad se refiere a la capacidad de entender el proceso de decisión del modelo de manera directa, sin necesidad de herramientas adicionales. En contraste, la explicabilidad implica proporcionar explicaciones adicionales para modelos complejos, a menudo descritos como caja negra. Esta distinción es importante al desarrollar herramientas XAI, ya que permite seleccionar el enfoque adecuado según el contexto y las necesidades de los usuarios.

5.2.3 Avances en Técnicas de Explicabilidad

El campo de la explicabilidad en inteligencia artificial (IA) presenta una amplia variedad de métodos y enfoques que van desde técnicas basadas en reglas y lógica, hasta métodos avanzados como el uso de Partes Prototípicas Profundas (PPs) y técnicas basadas en entropía. Esta diversidad refleja la complejidad y el dinamismo del campo, proporcionando diferentes

soluciones para abordar la necesidad de explicaciones claras y comprensibles de los modelos de IA.

Entre los métodos, SAMCNet (una red que mejora la interpretación de patrones espaciales) y Deep Prototypical-Parts (PPs) (técnicas que facilitan la identificación de características claves) sobresalen por sus mejoras en visualización y precisión, facilitando una interpretación más intuitiva y visual de las decisiones de los modelos. Por otro lado, técnicas como LIME y SHAP (métodos agnósticos del modelo que generan explicaciones basadas en ejemplos locales) continúan siendo populares debido a su flexibilidad y capacidad para generar explicaciones en una variedad de contextos y tipos de modelos, lo que se evidencia en su adopción en estudios recientes.

Sin embargo, algunos métodos, como se detalla en la columna de "Transformación de datos" de la Tabla 3, enfrentan limitaciones en términos de generalización. Estos métodos a menudo requieren perturbaciones o permutaciones del conjunto de datos original para generar explicaciones, lo que puede restringir su aplicabilidad en ciertos contextos donde se requiere integridad del conjunto de datos.

Los avances recientes, como las Explicaciones Lógicas Basadas en Entropía (que utilizan principios de la teoría de la información para proporcionar explicaciones más formales) y la Atribución Basada en Gradiente (que emplea derivadas para entender la influencia de cada característica en el resultado), han contribuido a mejorar la claridad y la robustez de las explicaciones para modelos complejos. Estas técnicas, que integran principios lógicos y matemáticos proporcionando explicaciones formales y cuantitativas, están ganando terreno en la investigación.

Además, a medida que crece el interés en técnicas más robustas, hay una tendencia hacia el desarrollo de métodos de explicabilidad que son generalizables a diferentes modelos de caja negra, como se describe en la columna "Generalizable". Esta capacidad de adaptación destaca la necesidad de enfoques universales para el futuro de la XAI.

La tabla también destaca la importancia de incluir ejemplos ilustrativos en los artículos revisados. Estos ejemplos no solo demuestran la aplicabilidad de los métodos, sino también su efectividad práctica. Este enfoque es esencial para mostrar cómo los métodos propuestos mejoran con respecto a los enfoques anteriores, como se detalla en la columna “Enfoques anteriores”, y para resaltar las principales mejoras o diferencias en comparación con las técnicas existentes.

Aunque métodos como SHAP y LIME siguen siendo ampliamente utilizados por su versatilidad, los enfoques innovadores como SAMCNet y los métodos basados en entropía ofrecen soluciones más especializadas que pueden ser más efectivas en aplicaciones

concretas. La comparación y evaluación continua de estos métodos ayudarán a identificar las mejores prácticas y a mejorar la explicabilidad de los modelos de IA en el futuro.

5.2.4 Desafíos

A pesar de los avances, persisten desafíos significativos en el campo de XAI. Crear técnicas de XAI que sean agnósticas al modelo y que puedan aplicarse de manera efectiva a una amplia variedad de modelos de IA sigue siendo un reto.

5.3 Implicaciones Prácticas y Políticas

5.3.1 Implicaciones Prácticas

La adopción de técnicas de XAI debería ser una práctica estándar en el desarrollo de sistemas de IA críticos. Los desarrolladores deberían integrar herramientas de explicabilidad desde las primeras etapas del diseño para garantizar que los sistemas sean transparentes y fomenten la confianza. Esto incluye la implementación de técnicas que permitan a los usuarios comprender cómo se toman las decisiones automatizadas y proporcionar mecanismos para revisar y cuestionar esas decisiones cuando sea necesario.

5.3.2 Implicaciones para la Política

A nivel regulatorio, se deben desarrollar políticas que promuevan la transparencia y la responsabilidad en el uso de IA. Las políticas deben establecer estándares claros para la explicabilidad y la transparencia, y fomentar prácticas que garanticen que los sistemas de IA operen de manera justa y respeten la privacidad de los individuos.

5.4 Trabajo Futuro

Este estudio proporciona una base para mejorar la comprensión y la transparencia en los modelos de IA, especialmente en áreas críticas como la medicina, las finanzas y la seguridad. Como trabajo futuro se propone centrarse en la implementación de alguna de las técnicas de XAI discutidas, evaluando su impacto en la comprensión y aceptación de la IA por parte de usuarios finales y expertos. Además, es importante explorar cómo adaptar estas técnicas a las necesidades específicas de diferentes sectores, asegurando que las decisiones basadas en IA sean transparentes, justas y respetuosas de la privacidad.

Anexo I

Análisis Detallado de 30 Trabajos Seleccionados sobre Avances Recientes en la Explicabilidad de Modelos de IA

Introducción

En este anexo se presentan los avances recientes en la explicabilidad de modelos de inteligencia artificial (IA) a través de una revisión de 30 trabajos seleccionados. Cada uno de ellos aborda diversos enfoques y técnicas que contribuyen a mejorar la comprensión y la transparencia de los modelos de IA, especialmente en contextos complejos y críticos.

La explicabilidad en modelos de IA garantiza que estos sistemas sean comprendidos y considerados confiables por los usuarios y responsables de la toma de decisiones. En este contexto, los papers seleccionados no solo presentan innovaciones en la explicación de modelos existentes, sino que también introducen nuevas técnicas que aportan mayor claridad y comprensión a los procesos.

Para cada uno de los papers revisados, se ofrece una breve descripción del enfoque y las contribuciones. Además, se presenta un cuadro resumen que evalúa aspectos considerados para su inclusión en este documento, asegurando que cumplen con la pregunta de investigación. Estos son:

- **Mejoras en la Explicabilidad**: Cómo el paper contribuye a mejorar la comprensión y transparencia de los modelos de IA.
- **Marco**: La relevancia del método o su comparación con otros enfoques en el contexto del estudio.
- **Comparación con Enfoques Anteriores**: Cómo el enfoque propuesto se compara con métodos anteriores en términos de efectividad y avance tecnológico.
- **Propuestas**: Las nuevas propuestas o técnicas introducidas por el paper para mejorar la explicabilidad.
- **Resultados Relevantes**: Los hallazgos clave y su impacto en la práctica de la IA explicable.

Este análisis proporciona una visión general de los avances en el campo y facilita la comparación entre diferentes enfoques y técnicas, destacando su contribución a la pregunta de investigación planteada.

El análisis detallado de estos papers se ha sintetizado en las Tablas 3, 4 y 5, que se encuentran en el cuerpo del documento. Estas tablas resumen los hallazgos y contribuciones principales, permitiendo una comprensión más profunda de cómo cada trabajo se relaciona con los avances recientes en la explicabilidad de modelos de IA.

A continuación, se presentan los papers seleccionados:

1. SAMCNet: Avances en la Clasificación Explicable con IA Espacial (Farhadloo et al., 2022) .

SAMCNet (Spatial-interaction Aware Multi-Category deep neural Network, SAMCNet por sus siglas en inglés) es un enfoque de clasificación con IA explicativa diseñado para aprender patrones espaciales en puntos de múltiples categorías y distinguir entre dos clases. La importancia de las configuraciones espaciales radica en su capacidad para generar hipótesis en la investigación de terapias para enfermedades como el cáncer, la investigación biomédica y la ecología microbiana, donde los patrones multicategóricos son comunes.

Este problema es complejo debido a la heterogeneidad de los patrones de puntos, sus interacciones espaciales estructurales y de orden superior, así como la necesidad de que el modelo distinga entre diversas interacciones espaciales. Aunque trabajos previos han utilizado medidas de asociación espacial y enfoques basados en gráficos vecinales, estos métodos suelen ser limitados por su sensibilidad a particionamientos espaciales fijos y su incapacidad para modelar relaciones espaciales complejas y direccionales.

Para abordar estas limitaciones, SAMCNet, incorpora capas de caracterización de marcos de referencia locales y priorización de pares de puntos.

1.1 Modelo para Clasificación Oncológica con IA

SAMCNet es un modelo propuesto para la clasificación en datos oncológicos basados en imágenes MxIF (Multiplexed Immunofluorescence Imaging, MxIF, por sus siglas en inglés). Se comparó con métodos tradicionales y arquitecturas DNN (Deep Neural Network, DNN por sus siglas en inglés),existentes como PointNet, DGCNN(Dynamic Graph Convolutional Neural Network, DGCNN por sus siglas en inglés**)** y SRNet (Spatial-Relationship Aware Neural Network, SRNet, por sus siglas en inglés), en tres tareas de clasificación:

- clasificación del margen del tumor,
- clasificación del núcleo del tumor y
- clasificación de enfermedades.

Las Relaciones Espaciales de Dos y Tres Vías se representan mediante hyperedges que conectan vértices de diferentes categorías, permitiendo modelar relaciones complejas y contextuales entre distintos tipos de puntos en los datos.

Para integrar la información espacial, se utiliza una función asimétrica como el promedio ponderado (average pooling). Esta función combina la información de todos los puntos vecinos alrededor del punto central v^i, reflejando la contribución diferencial de diferentes pares de categorías en la representación global de v^i.

1.2 Aplicación de la Tecnología MxIF en el Análisis Espacial de Microambientes Tumorales-Inmunes y Terapia con Inhibidores de Checkpoint Inmunitario.

El estudio se enfoca en el uso de la tecnología de inmunofluorescencia multiplexada (MxIF) para investigar los microambientes tumorales-inmunes (TME) y su aplicación en la terapia con inhibidores de checkpoint inmunitario (ICI). MxIF permite la identificación precisa de múltiples subtipos celulares y sus ubicaciones espaciales en tejidos fijados en formalina e incrustados en parafina, utilizando técnicas avanzadas de tinción y análisis de imágenes de células individuales. Es importante comprender las complejas interacciones espaciales entre células inmunes y cancerosas para mejorar la efectividad de ICI y avanzar en la investigación clínica en oncología.

1.3 Revisión de Métodos en el Análisis Espacial de Patrones de Puntos Multi-Categóricos: Enfoques Basados en Datos y Redes Neuronales Profundas

En el campo del análisis espacial de patrones de puntos multi-categóricos, se han explorado dos enfoques principales:

Cuantificación espacial basada en datos: Este enfoque se centra en métodos que emplean medidas de asociación espacial para comprender las interacciones entre diferentes categorías de puntos. Se utilizan técnicas como la correlación de Pearson, Cross-k, G-cross, e índice de participación (Shekhar & Huang, 2001). Como ejemplo, la correlación de Pearson ha sido aplicada para estudiar la asociación espacial entre células tumorales e inmunes en imágenes digitalizadas de cáncer de mama (Maley et al. (2015)).

Características construidas por máquinas utilizando redes neuronales profundas (DNN): Este enfoque utiliza redes neuronales profundas para modelar relaciones espaciales entre puntos de diferentes categorías. Por ejemplo SRNet, es una red neuronal diseñada para ser sensible a relaciones espaciales, aunque está limitada a relaciones binarias y asume igual importancia entre los pares de categorías binarias (Li et al. (2021)). SRNet utiliza una distancia de vecindario fija para construir el grafo de entrada, lo cual puede ser restrictivo en entornos con vecindarios de diferentes tamaños.

El texto también señala una crítica hacia la mayoría de las DNN basadas en características revisadas en estudios recientes de patología computacional, que utilizan imágenes de rejillas regulares como entrada. Esto limita su capacidad para manejar estructuras geométricas simples pero importantes como los patrones de puntos.

1.4 Trabajo Propuesto: SAMCNet

El objetivo principal de la arquitectura propuesta de red neuronal es aprender relaciones espaciales de N-vías en un patrón de puntos de múltiples categorías. La principal diferencia entre SAMCNet y las medidas tradicionales de interés de asociación basadas en datos es LRFC, que permite que puntos categóricos pertenecientes a diferentes distribuciones (por ejemplo, agrupamiento versus distribución uniforme) se presenten a través de una representación multi-escala, superando la ineficiencia de métodos de una sola escala como los núcleos de función de base radial o la discretización. Además, SAMCNet difiere de la arquitectura de redes neuronales profundas (DNN) SRNet en dos aspectos:

- Incorpora una subred de priorización de pares de puntos, que aprende la importancia de los pares de puntos en relaciones espaciales de N-vías basadas en sus atributos categóricos.
- La conectividad de los nodos en un grafo localmente conectado permite que SAMCNet aprenda patrones espaciales relevantes de alto orden basados en los *k* vecinos más cercanos de entrada y la elección de agregación.

1.5 Caracterización del Marco de Referencia Local

Dado un conjunto de puntos $P = \{pi = (ci, fi)|p1, ..., pn\}$, donde $ci = (xi, yi)$ son las coordenadas espaciales y *fi* es el atributo categórico, calculamos un grafo dirigido $G = (V, E)$, donde V y E son los vértices y aristas. Construimos G como el grafo de k-vecinos más cercanos de cada punto $ci \in \mathrm{IR}F$, donde $E = V \times K$.
Tener en cuenta que el grafo de vecindario de cada capa consecutiva en SAMCNet depende de la salida de la capa anterior, que se actualiza dinámicamente en función de la dimensión F, que representa la dimensionalidad de características de la red (Wang et al., 2019) . El atributo categórico *fi* asociado con cada *pi* se conserva en toda la red mediante una conexión de omisión para los cálculos de subred de priorización de pares de puntos (Mai et al. (2020)).
SAMCNet está diseñado para capturar las relaciones espaciales complejas en datos biomédicos, como los obtenidos mediante inmunofluorescencia multiplex (MxIF) en oncología. La arquitectura consta de varios componentes claves que trabajan juntos para clasificar los datos basándose en sus patrones espaciales.

El siguiente paso es utilizar la caracterización del marco de referencia local (LRFC) para modelar la distribución de un punto y su vecino utilizando solo coordenadas espaciales. Esta técnica permite modelar la distancia relativa entre un punto dado *ci* con respecto a sus puntos más cercanos *cj*, donde $1 \leq j \leq k$, en una arista correspondiente *e'ij*. La intuición detrás de LRFC es que las coordenadas espaciales son indicadores ilustrativos de ubicación; el uso de técnicas de discretización o redes neuronales feedforward es insuficiente para capturar la distribución espacial debido a la falta de descomposición de características entre atributos espaciales y categóricos (Abbott & Callaway, 2014).

Inspirados en una representación periódica multi-escala de celdas de cuadrícula en mamíferos (Abbott & Callaway, 2014) y en una representación vectorial de auto-posición (Gao, Xie, Zhu, & Wu, 2018), Mai et al. (Gao, Xie, Zhu, & Wu, 2018) propusieron un método de incrustación multi-escala, es decir, la codificación posicional *PE*, que utiliza funciones seno y coseno de diferentes frecuencias para representar posiciones en el espacio. Se adopta esta idea en la red de la siguiente manera.

Dado un punto *ci* en un espacio 2D estudiado, $e[ci] = EdgeConv(PEs(ci))$, donde $PE(ci)$ es una representación multi-escala *sj*, $1 \le j \le s$, para capturar la distribución de patrones de puntos de múltiples categorías mixtas. La formulación general de la caracterización del marco de referencia local (LRFC) es la siguiente:

$$PE_s(c_i) = [PE_{s,1}(c_i); \ldots; PE_{s,s}(c_i)], \quad (1)$$

$$PE_{s,j}(c_i) = [\cos(\frac{\langle c_i, a_j \rangle}{\lambda_{min} \cdot g^{s/(S-1)}}); \sin(\frac{\langle c_i, a_j \rangle}{\lambda_{min} \cdot g^{s/(S-1)}})], \quad (2)$$

$$\forall j = 1, 2, 3,$$

$$\text{where } a_1 = [1, 0]^T, a_2 = [-1/2, \sqrt{3}/2]^T, \text{ and } a_3 = [-1/2, -\sqrt{3}/2]^T$$

para todo $j = 1, 2, 3$, donde $a1 = [1, 0]T$, $a2 = [-1/2, \sqrt{3}/2]T$, y $a3 = [-1/2, -\sqrt{3}/2]T$ son vectores unitarios, los ángulos entre cada par de vectores son $2\pi/3$; λmin y λmax son las escalas mínima y máxima de la cuadrícula; y $g = \lambda max/\lambda min$. Se define la incrustación a lo largo de cada arista como la distancia entre el punto central *ci* y sus k-vecinos más cercanos *cj*, $|PE(ci) - PE(cj)|$, donde $1 \le j \le k$.

1.6 Medición a escala Local y Global

En la arquitectura SAMCNet, la operación EdgeConv se define como la característica de borde $eij = h\Theta(ci, cj)$, donde $h\Theta$ es una función no lineal con parámetros aprendibles Θ que mapea de $RF \times RF$ a RF'. Posteriormente, se aplica una operación asimétrica (por ejemplo, P o Max) para agregar información a lo largo de todas las características de borde que rodean al nodo central *ci*. La elección de $h\Theta$ es crucial para definir EdgeConv, cómo utilizar el producto punto entre un conjunto de filtros $\Theta = \{\theta 1, \ldots, \theta M\}$ y píxeles de imagen *xj* en una cuadrícula regular, y luego agregar la información usando P, lo que resulta en una convolución estándar. Para una discusión detallada de las diferentes formas de $h\Theta$, (Wang et al. (2019)).

Se ha adaptado la operación EdgeConv de DGCNN (Wang et al. (2019)) en la red para aprender tanto la estructura global de forma, capturada por las coordenadas centrales *ci*, como

la información local del vecindario, capturada por |*ci* − *cj*|. La formulación general es la siguiente:

$$e''_{ij} = leakyrelu(\theta_m.|PE(c_i) - PE(c_j)| + \phi_m.c_i),$$

donde θm y ϕm son parámetros aprendibles para la información local y global, respectivamente, y *PE* es la incrustación posicional que representa las distancias relativas a lo largo de cada borde comenzando en *ci*.

1.7 Subred de Priorización de Pares de Puntos

Hasta ahora, se ha construido el grafo y definido las incrustaciones de borde en términos estrictamente de características espaciales. Si seguimos arquitecturas DNN basadas en grafos de patrones de puntos existentes, simplemente se concatenan las características categóricas en el espacio de características incrustadas. Sin embargo, de esta manera no se aprendería la importancia de las interacciones entre vértices de características categóricas *fi* y *fj* ∈ N*i*. Como resultado, el modelo estaría limitado a aprender características individuales de categorías.

En lugar de eso, el clasificador debe aprender cómo ponderar correctamente las diversas asociaciones de pares de puntos como un sesgo inductivo más fuerte. Con este fin, se propone una capa de priorización de pares de puntos para aprender la importancia (es decir, la fuerza) de la relación espacial entre diferentes pares de categorías, seguida de una capa de agrupación promedio para ponderar los diferentes subconjuntos en consecuencia. En conjunto, esta capa es análoga a una función de agrupación promedio ponderada, donde los pesos corresponden a la importancia de la interacción categórica.

El input a esta capa es una incrustación de borde *e'ij*, que es la salida de la capa EdgeConv. En la capa de priorización, primero derivamos *eˆij*, una incrustación de borde aumentada por la fuerza de la asociación par a par categórica:

eˆ_{ij} = *a^®_{fi fj}* \cdot *W* \cdot *e"_{ij}*,

donde *W* es una transformación lineal aprendible en la incrustación original para ayudar a la expresividad de la priorización, y *a^®fi fj* es nuestro vector de peso de asociación par a par aprendido para características de pares de puntos categóricos (*fi*, *fj*).

En esta formulación, se incluyó *fj* ∈ N*i*, donde *a^®fi fj* es un auto-peso aprendido basado solo en la característica categórica de *vi*. También se asume que las interacciones son invariantes con respecto al orden de las categorías; por ejemplo, αC1C2 ≡ αC2C1.

Similar a otras capas de priorización (es decir, atención), se aplica una función de activación LeakyReLU (LR), seguida de una función softmax, resultando en la asociación normalizada par a par:

$$\alpha_{f_i f_j} = \frac{\exp(\mathrm{LR}(\hat{e}_{ij}))}{\sum_{k \in \mathcal{N}_i} \exp(\mathrm{LR}(\hat{e}_{ij}))},$$

donde *αfi fj* es la asociación par a par categórica aprendida para cada vecino, de modo que *fj* ∈ IR*k*. Con este coeficiente de atención normalizado, *αfi fj*, podemos calcular la agrupación promedio ponderada y producir la incrustación final del vértice:

$$\hat{v}_i = \sigma\left(\frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \alpha_{f_i f_j} W e''_{ij}\right).$$

Esta formulación se puede extender a *K*, siguiendo otras redes de priorización como GAT (Veličković et al. (2017)), donde cada cabeza aprende una asociación par a par categórica separada *αkfi fj* y un peso de transformación lineal *Wk*, seguido de una operación de agregación (AGG) sobre las salidas de las diferentes cabezas:

donde σ es una función de activación no lineal como LeakyRelu y la operación de agregación puede tomar la forma de un promedio o una concatenación.

Dado que esta capa preserva la identidad del vértice central, también se puede extender a múltiples capas de la red manteniendo las características categóricas de los vértices entre capas con una conexión de omisión. Esto se logra añadiendo la priorización de pares de puntos después de la operación EdgeConv de cada capa. En el contexto del aprendizaje jerárquico de características, nuestra red es efectivamente capaz de aprender la importancia de las interacciones categóricas N-way en un espacio de características jerárquico.

Finalmente, señalamos que la elección de la agregación no se limita a la agrupación promedio; por ejemplo, se puede elegir un gran número de vecinos más cercanos al construir el grafo, mientras se toma solo un subconjunto superior-*k′* de las características más altas para agrupar, para filtrar un número abrumador de interacciones débiles. La agrupación máxima se puede demostrar como un caso especial de este concepto, donde solo se selecciona la parte superior-1 de las características de un vecino.

1.8 Resultado de la experimentación

SAMCNet es un modelo propuesto para la clasificación en datos oncológicos basados en imágenes MxIF. Se comparó con métodos tradicionales y arquitecturas DNN existentes como PointNet, DGCNN y SRNet en tres tareas de clasificación: clasificación del margen del tumor, clasificación del núcleo del tumor y clasificación de enfermedades.

Resultados Comparativos: SAMCNet mostró una mejora significativa en precisión, recall, F1-score y exactitud (ACC) en comparación con los métodos tradicionales y las DNN existentes. En particular, superó a SRNet en un 7.0% y un 17.0% en las clasificaciones de margen tumoral y enfermedades, respectivamente.

Complejidad Temporal: SAMCNet demostró ser mucho más rápido que SRNet en las tres tareas de clasificación, con tiempos de inferencia reducidos de 3 a 10 veces.

Análisis de Sensibilidad: Se evaluó cómo diferentes componentes dentro de SAMCNet afectan su rendimiento. Se encontró que el uso de una subred de priorización de pares de puntos mejora significativamente el filtrado de interacciones débiles y que la caracterización del marco de referencia local (LRFC) es crucial para representar distancias relativas y distribuciones de manera efectiva.

Impacto de la Subred de Priorización: SAMCNet permite la interpretación de sus decisiones al medir la importancia de las asociaciones espaciales aprendidas entre diferentes tipos de células. Esto ayuda a distinguir entre patrones de puntos de diferentes clases (por ejemplo, respondedores y no respondedores) en el tumor-core.

Relaciones Espaciales N-way más Relevantes: El modelo identificó las interacciones espaciales de alto orden más significativas, como las asociaciones entre células tumorales, macrófagos y vasculatura, proporcionando insights críticos para el análisis oncológico.

En resumen, SAMCNet no solo superó a métodos tradicionales y DNN existentes en términos de precisión y velocidad, sino que también mejoró la interpretabilidad al revelar las interacciones espaciales clave que influyen en la clasificación de patrones celulares complejos en datos MxIF oncológicos.

1.9 Respuesta a la pregunta de investigación.

Mejoras en la Explicabilidad

Modelos de Redes Neuronales Conscientes del Espacio: Uno de los avances más significativos es el desarrollo de modelos como el Spatial-Relationship Aware Neural Network (SRNet) y, más recientemente, SAMCNet. Estos modelos están diseñados para

capturar las relaciones espaciales entre puntos de datos, lo cual es crucial en aplicaciones como la clasificación de imágenes de tejidos en oncología .

Marco SAMCNet: El marco SAMCNet introduce una arquitectura de red neuronal que aprende relaciones espaciales N-arias en patrones de puntos de múltiples categorías. Esto se diferencia de los métodos tradicionales, que solo consideran relaciones binarias y usan métodos de cuantificación espacial que no capturan adecuadamente las interacciones espaciales complejas .

Comparación con Enfoques Anteriores:

SAMCNet se diferencia de la arquitectura de red neuronal profunda de última generación SRNet en dos formas principales:

1. Sub-Red de Priorización de Pares de Puntos:
    - SAMCNet incluye una sub-red que prioriza los pares de puntos. Esta sub-red aprende la importancia de los pares de puntos en las relaciones espaciales de múltiples vías (N-way) basándose en sus atributos categóricos. Esto permite que la red neuronal identifique qué pares de puntos son más relevantes para la tarea de clasificación.
2. Conectividad en un Grafo Localmente Conectado:
    - La conectividad de nodos en un grafo localmente conectado permite que SAMCNet aprenda patrones espaciales de orden superior relevantes. Utiliza los *k* vecinos más cercanos y una elección de agregación para aprender estos patrones. Esto significa que la red puede captar interacciones más complejas entre los puntos categóricos en el espacio.

1.10 Impacto de las Mejoras.

Propuestas y Taxonomías: SAMCNet proporciona un marco para clasificar datos mediante la identificación de las características explicativas más discriminativas basadas en su disposición espacial. Esto se ejemplifica en el uso de un índice de participación para distinguir clases de respuesta a tratamientos oncológicos basados en la disposición de células específicas (p. ej., células B y células T reguladoras) .

Resultados Relevantes: La implementación de SAMCNet ha permitido una mejor comprensión de las interacciones espaciales críticas en el microambiente tumoral. Esto puede mejorar la precisión del diagnóstico clínico y proporcionar nuevos conocimientos para la terapia del cáncer, destacando la importancia de los análisis espaciales informados en las respuestas de tratamiento .

1.11 Aspectos Destacados de SAMCNet en Clasificación Explicable de Datos Oncológicos

A continuación se incluye una comparación y evaluación del Modelo SAMCNet que se presenta en la Tabla 8.

Tabla 8. Comparación y evaluación de explicaciones basadas en el modelo SAMCNet para clasificación de datos oncológicos.

Aspecto	Descripción
Mejoras en la Explicabilidad	Modelos como SRNet y SAMCNet capturan relaciones espaciales en datos, importante para la clasificación de imágenes de tejidos en oncología.
Marco SAMCNet	Introduce una arquitectura para aprender relaciones espaciales N-arias en patrones de múltiples categorías, superando métodos tradicionales de relaciones binarias.
Comparación con Enfoques Anteriores	Redes Neuronales Profundas (DNN) existentes:**PointNet; DGCNN** (Dynamic Graph Convolutional Neural Network);**SRNet** (Spatial Relation Network).
Propuestas	SAMCNet propone una arquitectura de red neuronal para entender la configuración espacial de patrones de puntos multi-categóricos en datos de oncología MxIF, enfocándose en las interacciones entre células tumorales como macrofagos asociados al tumor (TAMs) y neutrofilos asociados al tumor (TANs).
Resultados Relevantes	Mejora la comprensión de interacciones espaciales en el microambiente tumoral, importante para diagnósticos precisos y terapia del cáncer.

1.12 Avances en Explicabilidad de Modelos 'Caja Negra' y su Impacto en Aplicaciones Clínicas: SAMCNet en Oncología.

Los avances recientes en la explicabilidad de modelos de machine learning 'caja negra' han llevado a enfoques más sofisticados como SAMCNet, que superan las limitaciones de

métodos anteriores al capturar relaciones espaciales complejas. Estos desarrollos no solo mejoran la capacidad de los modelos para ser explicables y precisos, sino que también tienen un impacto significativo en aplicaciones clínicas, particularmente en la oncología, al proporcionar insights detallados y clínicamente relevantes sobre las interacciones celulares.

2. Explicaciones Basadas en Subobjetivos para Sistemas de Soporte de Decisión Inteligente No Confiables (Das, Kim, & Chernova, 2023).

El estudio aborda la necesidad de mejorar la explicabilidad en sistemas de soporte de decisión inteligente (IDS) basados en planificación, especialmente cuando estos sistemas pueden ser ocasionalmente subóptimos o poco confiables (Das et al., 2023). Se introduce el concepto de explicaciones basadas en submetas (ES B, Subgoal-Based Explanations, por sus siglas en inglés), inspiradas en la psicología humana que sugiere que descomponer tareas complejas en submetas más simples puede mejorar la comprensión y el desempeño del usuario (Simon, 1975).

2.1 Métodos

Se realizó un estudio experimental con 105 participantes reclutados de Amazon Mechanical Turk, divididos en diferentes condiciones de estudio para evaluar el impacto de las explicaciones ES B en comparación con otras formas de explicación (Causal-Link-Chain Explanations, EC L C, por sus siglas en inglés, y Action Recommendations from IDS, aIDS, por sus siglas en inglés). Los participantes jugaron un juego de simulación de planificación de restaurantes donde debían completar entregas de comidas dentro de tiempos específicos.

Desglose de Participantes:

- Reclutamiento: Los 105 participantes fueron reclutados a través de Amazon Mechanical Turk, asegurando una muestra diversa en términos de edad y antecedentes.
- Condiciones del Estudio: Los participantes fueron divididos en siete condiciones experimentales, incluyendo un grupo de control sin ayuda del IDS y varios grupos con diferentes tipos de explicaciones (ES B, EC L C, aIDS) bajo condiciones óptimas y subóptimas del IDS.

2.2 Rendimiento de la Tarea y Costo del Plan del Usuario (UPC)

Los participantes que recibieron explicaciones basadas en submetas (ES B) mostraron significativamente menores costos de planificación (UPC) en comparación con aquellos que recibieron solo recomendaciones de acción (Action Recommendations from IDS, *aIDS* por sus siglas en inglés) o explicaciones de cadena de vínculo causal (Causal-Link-Chain Explanations, EC L C, por sus siglas en inglés) (Das et al., 2023). Esta mejora fue consistente

tanto en condiciones óptimas como subóptimas del sistema de soporte de decisión inteligente (IDS).

Análisis de Datos.

Metodología Estadística: Se utilizaron pruebas t y análisis de varianza (ANOVA) para determinar la significancia estadística de los resultados, asegurando la validez y confiabilidad de los hallazgos.

2.3 Capacidad para Distinguir Recomendaciones Óptimas vs. Subóptimas

Las explicaciones basadas en submetas (ES B) permitieron a los usuarios identificar y evitar recomendaciones menos efectivas del sistema de soporte de decisión inteligente (IDS) mejor que otras formas de explicación, como demostró el porcentaje de acciones subóptimas evitadas (SAA%, Suboptimal Action Avoidance percentage por sus siglas en inglés) (Das et al., 2023).

Comparación con Otros Métodos.

Explicaciones Contrastivas: Las explicaciones contrastivas explican por qué se eligió una acción en lugar de otra, pero pueden ser menos intuitivas que las explicaciones ES B.

Explicaciones de Cadena de Vínculo Causal: Aunque útiles para usuarios con experiencia técnica, estas explicaciones pueden ser demasiado complejas para usuarios novatos.

2.4 Preferencia del Usuario por las Explicaciones ES B

Hubo una clara preferencia de los usuarios por las explicaciones ES B sobre EC L C y *aIDS*, indicando que estas explicaciones son más intuitivas y útiles para comprender las recomendaciones del IDS (Das et al., 2023).

Resultados de Preferencia del Usuario.

Se utilizaron encuestas para medir la preferencia de los participantes, mostrando una preferencia significativa por ES B en términos de claridad y utilidad.

2.5 Robustez en Escenarios de Falla del IDS

Incluso en la etapa de Evaluación donde el IDS no estaba disponible, los participantes que previamente recibieron explicaciones ES B demostraron un menor UPC en comparación con otras condiciones de estudio, destacando la utilidad a largo plazo de las explicaciones ES B (Das et al., 2023).

Aplicabilidad en Diferentes Contextos:

- Escenarios de Falla: Las explicaciones ES B demostraron ser efectivas incluso cuando el IDS fallaba, sugiriendo su aplicabilidad en contextos donde la confiabilidad del sistema es variable.

2.6 Respuesta a la pregunta de investigación

Mejora del Rendimiento del Usuario: Las explicaciones basadas en submetas (ES B) mejoran significativamente el rendimiento del usuario, demostrado por menores costos de planificación del usuario (UPC) en comparación con otras formas de explicación como recomendaciones de acción (*aIDS*) o explicaciones de cadena de vínculo causal (EC L C) (Das et al., 2023).

Capacidad para discernir Recomendaciones Óptimas vs. Subóptimas: ES B ayuda a los usuarios a identificar y evitar recomendaciones subóptimas del IDS de manera más efectiva que otras formas de explicación, como se refleja en un mayor porcentaje de Evitación de Acciones Subóptimas (SAA%) (Das et al., 2023).

Preferencia del Usuario y Comprensión Mejorada: Los usuarios prefieren significativamente las explicaciones ES B sobre EC L C y *aIDS*, indicando que estas explicaciones son más intuitivas y útiles para comprender las recomendaciones del IDS (Das et al., 2023).

Robustez en Escenarios de Falla del IDS: Incluso cuando el IDS no está disponible, los usuarios que previamente recibieron explicaciones ES B muestran un rendimiento continuamente mejorado, destacando la utilidad a largo plazo de estas explicaciones (Das et al., 2023).

2.7 Aspectos y Resultados de la Introducción de Explicaciones Basadas en Submetas (ES B) en Sistemas de Soporte de Decisión Inteligente (IDS).

A continuación se incluye una comparación y evaluación del Modelo de Explicaciones basada en submetas que se presenta en la Tabla 10.

Tabla 10. Comparación y evaluación de explicaciones basadas en submetas en sistemas de soporte de decisión inteligente N(IDs)

Aspecto	Descripción
Mejoras en la Explicabilidad	Introducción de explicaciones basadas en submetas (ES B) en sistemas de soporte de decisión inteligente (IDS), permitiendo que las recomendaciones estén fundamentadas en submetas.
Marco ES B (Subgoal-Based Explanations)	Basado en psicología cognitiva, los humanos tienden naturalmente a dividir

	tareas complejas en submetas más manejables, lo cual ha inspirado el diseño de ES B.
Comparación con Enfoques Anteriores	Los enfoques anteriores incluyen recomendaciones de acción (*aIDS*, Action Recommendations from IDS) y explicaciones de cadena de vínculo causal (EC L C, Causal-Link-Chain Explanations), que son menos efectivos para lograr una comprensión profunda de las recomendaciones del IDS.
Impacto de las Mejoras	Mejora la capacidad de los usuarios para identificar y evitar recomendaciones subóptimas, y mejora el rendimiento general en tareas complejas.
Propuestas y Taxonomías	Propone integrar técnicas de identificación automática de submetas para mejorar la eficacia de subgoal-based explanations (ES B) en plan-based Intelligent Decision Support Systems (IDS).
Resultados Relevantes	Mejoras en el rendimiento de la tarea: Los usuarios que recibieron explicaciones basadas en submetas (ES B) mostraron un costo de planificación significativamente menor (User Plan Cost, UPC) en comparación con aquellos que recibieron recomendaciones de acción (*aIDS*) o explicaciones de cadena de vínculo causal (EC L C). Esto se observó tanto en la etapa con soporte del Sistema de Soporte de Decisiones Inteligentes (IDS) como en la etapa de evaluación sin soporte del IDS. Preferencia del usuario por ES B: En un experimento adicional, los usuarios mostraron una preferencia significativa por las explicaciones basadas en submetas (ES B) sobre las recomendaciones de acción (*aIDS*) y las explicaciones de cadena de vínculo causal (EC L C) para comprender la próxima acción del chef en el juego de planificación del restaurante.

	Conformidad con la acción óptima (OAC%) y capacidad para evitar acciones subóptimas (SAA%): Durante la etapa con soporte del IDS, los usuarios que recibieron ES B mostraron una alta conformidad con las acciones sugeridas y una mejor capacidad para evitar acciones subóptimas en comparación con los que recibieron *aIDS* o EC L C.

2.8 Métodos de explicabilidad utilizados en sistemas de soporte de decisión inteligente (IDS)

Recomendaciones de Acción (Action Recommendations from IDS, *aIDS*): Este método se centra en proporcionar al usuario recomendaciones directas sobre las acciones a tomar. Sin embargo, carece de explicaciones detalladas sobre el razonamiento detrás de las recomendaciones, lo que puede llevar a una menor comprensión y confianza por parte del usuario.

Explicaciones de Cadena de Vínculo Causal (Causal-Link-Chain Explanations, EC L C): Este enfoque intenta proporcionar una explicación detallada de las recomendaciones del IDS mediante la presentación de la cadena de vínculos causales que llevaron a la recomendación. Si bien esto puede ser útil para usuarios avanzados, puede resultar confuso y abrumador para usuarios no expertos.

Explicaciones Basadas en Submetas (Subgoal-Based Explanations, ES B): Las explicaciones basadas en submetas descomponen la recomendación del IDS en submetas más manejables y comprensibles. Este enfoque, basado en la psicología cognitiva, mejora la comprensión del usuario sobre el por qué de las recomendaciones, facilitando una mejor toma de decisiones y una mayor capacidad para evitar recomendaciones subóptimas.

Impacto de las explicaciones basadas en submetas en el rendimiento del usuario:

- Menor Costo de Planificación del Usuario (UPC): Los participantes que recibieron explicaciones ES B lograron menores costos de planificación, tanto en condiciones óptimas como subóptimas del IDS.
- Mayor Conformidad con Acciones Óptimas (OAC%): Las explicaciones ES B ayudaron a los usuarios a seguir más de cerca las recomendaciones óptimas del IDS.
- Mejor Capacidad para Evitar Acciones Subóptimas (SAA%): Los usuarios con explicaciones ES B pudieron identificar y evitar acciones subóptimas más efectivamente.

2.9 Avances en la explicabilidad basada en submetas.

Las explicaciones basadas en submetas (ES B) representan una mejora significativa en la capacidad de los sistemas de soporte de decisión inteligente (IDS) para comunicar recomendaciones de manera efectiva y comprensible. Este enfoque no solo mejora el rendimiento del usuario en tareas complejas, sino que también aumenta la confianza y preferencia del usuario por el sistema.

3. LCNN: Arquitectura de CNN Ligera para la Identificación de Características de Defectos de Software Usando IA Explicable (Begum, Alam, Islam, & Hossain, 2024).

El paper "LCNN: Lightweight CNN Architecture for Software Defect Feature Identification Using Explainable AI" aborda el desafío de mejorar la interpretabilidad de los modelos de machine learning (ML), específicamente de redes neuronales convolucionales (CNN), aplicadas a la identificación de defectos en software. Se presenta una arquitectura de CNN combinada con técnicas de inteligencia artificial explicable (XAI) para proporcionar un entendimiento claro y detallado de las predicciones del modelo.

3.1 Avances en la identificación de defectos de software

A continuación se mencionan avances en la identificación de defectos de software, específicamente centrados en el uso de redes neuronales convolucionales (CNN) y técnicas de inteligencia artificial explicable (XAI).

1. Uso de CNNs y Arquitecturas Personalizadas: Hay un enfoque creciente en el desarrollo de arquitecturas ligeras y adaptadas de CNN para mejorar la eficiencia y precisión en la identificación de defectos en sistemas de software. Estas arquitecturas están diseñadas específicamente para manejar las complejidades del código de software y garantizar transparencia en el proceso de identificación.
2. Integración de Técnicas de XAI: Se destaca la importancia de integrar técnicas de XAI en estos modelos para mejorar la interpretabilidad. Esto implica proporcionar insights sobre el proceso de toma de decisiones de la red neuronal, lo que ayuda a entender las causas subyacentes de los defectos de software identificados.
3. Estudios y Modelos Propuestos:
    - **Tong et al.,(2018):** Propusieron un enfoque para resolver el problema del desequilibrio de clases en la predicción de defectos de software utilizando representaciones profundas y un ensamble de dos etapas. Sin embargo, no abordaron la predicción de defectos entre proyectos diferentes (cross-project).
    - **Zhu et al. (2020):** Introdujeron un modelo de predicción de defectos denominado DAECNN-JDP, que combina técnicas de autoencoders para

eliminar ruido con CNNs para predecir defectos justo a tiempo. Este modelo superó a 11 modelos de referencia en proyectos de código abierto, aunque no se evaluaron en proyectos comerciales.

- **Qiu et al. (2019):** Presentaron un método que utiliza modelos CNN de transferencia para extraer características semánticas transferibles en tareas de predicción de defectos entre proyectos diferentes (CPDP).

4. Consideraciones Adicionales:
   - Aprendizaje Sensible al Costo: Se discute la efectividad del aprendizaje sensible al costo para abordar el desequilibrio en los costos asociados con la clasificación incorrecta de módulos defectuosos y no defectuosos.
   - Modelos Híbridos y Mejorados: Se mencionan modelos híbridos que combinan LSTM bidireccionales con CNN y marcos de análisis de dependencia semántica para mejorar la precisión en la predicción de problemas de software.

3.2 Metodología usada.

La metodología propuesta incluye el diseño de una arquitectura de CNN ligera y la integración de dos técnicas de XAI: LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive Explanations). A continuación se detallan las técnicas y fórmulas utilizadas:

1. Arquitectura LCNN:
   - Se diseñó una arquitectura CNN simplificada para reducir la complejidad y mejorar la velocidad de procesamiento sin sacrificar la precisión.
   - Ecuación de la convolución:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$

   - El modelo utiliza varias capas convolucionales seguidas de capas de pooling y fully connected.

2. Explicabilidad con LIME y SHAP:
   - LIME (Local Interpretable Model-agnostic Explanations): Proporciona explicaciones locales generando aproximaciones lineales del modelo en torno a una predicción específica.

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

donde π_x es una medida de proximidad alrededor de x, L es una función de pérdida que mide la precisión del modelo interpretable g respecto al modelo original f, y Ω es una penalización por complejidad del modelo g.

- SHAP(Basado en los valores de Shapley): proporciona explicaciones consistentes y globales sobre la importancia de las características.

- **Fórmula de los valores de Shapley:**

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

donde ϕ_i es el valor de Shapley para la característica i, N es el conjunto de todas las características, S es un subconjunto de N que no incluye i, y v es la función de valor que evalúa el rendimiento del modelo.

3.3 Principales Resultados sobre la arquitectura LCNN

- Eficiencia y Precisión: La arquitectura LCNN mostró una mejora en la eficiencia y precisión comparada con modelos tradicionales.
- Explicabilidad:.LIME y SHAP son técnicas valiosas para la explicabilidad, ofrecen diferentes enfoques sobre el proceso de toma de decisiones de la arquitectura LCNN, proporcionando información complementaria.
- Aplicaciones Prácticas: Se presentan estudios de casos que demuestran la aplicabilidad del modelo en diferentes dominios de software.

3.4 Enfoque propuesto.

El enfoque propuesto consta de cinco etapas, que componen la arquitectura del enfoque sugerido. La primera fase consiste en la utilización de conjuntos de datos CM1 de la NASA, técnicas de preprocesamiento y división de datos para 1D-CNN y 2D-CNN. En la segunda fase, experimentamos con 1D-CNN y luego, en la tercera fase, experimentamos con 2D-CNN utilizando datos procesados. En la cuarta etapa, comparamos 1D-CNN y 2D-CNN para

seleccionar la mejor técnica CNN para la predicción de defectos de software. Finalmente, mostramos cómo se utilizan LIME y SHAP para encontrar la explicación adecuada y visualizar la causa raíz de los defectos de software.

A continuación se describen cada una de las fases con sus características:

1. Primera Fase: Preparación y Preprocesamiento del Conjunto de Datos
    - **Conjunto de Datos:** Utilizamos el conjunto de datos CM1 de la base de datos PROMISE, que incluye 22 características para la predicción de defectos en el software, con un total de 6992 instancias (3455 con fallos y 3537 sin fallos).
    - **Preprocesamiento:** El conjunto de datos se somete a codificación de etiquetas (LE) y escalado estándar. La codificación de etiquetas asigna un número a cada categoría en una variable categórica para facilitar su manejo. El escalado estándar normaliza los datos para que su distribución tenga una media de cero y una desviación estándar de uno, usando la siguiente fórmula:

$$z_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

where μ = Mean and σ = Standard Deviation.

      Finalmente, dividimos el conjunto de datos en un 80% para entrenamiento y un 20% para prueba.

2. Segunda Fase: 1D-CNN
    - **1D-CNN:** Utilizamos redes neuronales convolucionales unidimensionales (1D-CNN) para trabajar con datos secuenciales. Las capas convolucionales extraen características jerárquicas automáticamente de los datos de entrada. La arquitectura de 1D-CNN incluye capas convolucionales unidimensionales, capas de pooling y capas completamente conectadas. Usamos la función de activación ELU (Exponential Linear Unit) para clasificar los fallos de software.
        - **Capa Convolucional 1D:** Aplica operaciones de convolución a lo largo de un solo eje, extrayendo características y patrones de los datos secuenciales.
        - **Función de Activación:** La ELU se utiliza para introducir no linealidad y mejorar la precisión de los modelos de aprendizaje.
        - **Stride:** Determina el tamaño del paso de la convolución a lo largo de la señal de entrada.

        - **Capa de Pooling:** Reduce la dimensionalidad y la carga computacional seleccionando el valor máximo dentro de una ventana (Max Pooling).
        - **Capa de Dropout y Capa Flatten:** La capa de dropout desactiva neuronas aleatoriamente durante el entrenamiento para prevenir el sobreajuste, y la capa flatten convierte matrices multidimensionales en un vector unidimensional para la conexión con capas completamente conectadas.
3. Tercera Fase: 2D-CNN
    - **2D-CNN:** Proponemos el uso de redes neuronales convolucionales bidimensionales (2D-CNN) para la predicción de defectos en el software utilizando datos tabulares transformados en representaciones 2D, como matrices, que permitan aplicar capas de CNN para la extracción de características.
        - **Transformación de Datos:** Los datos tabulares se transforman en un formato 2D similar a una imagen, representando relaciones entre métricas de software.
        - **Normalización e Ingeniería de Características:** Para mejorar la capacidad de la red para discernir patrones.
        - **Diseño de la Arquitectura 2D-CNN:** Incluye capas convolucionales y de pooling para capturar dependencias de características locales y globales.
        - **Entrenamiento y Validación:** Dividimos el conjunto de datos en entrenamiento, validación y prueba. Entrenamos el modelo ajustando hiper parámetros y validamos el rendimiento utilizando métricas como precisión, recall, F1 score y exactitud.
4. Cuarta Fase: Comparación de Modelos 1D-CNN y 2D-CNN
    - **Métricas de Evaluación:** Utilizamos varias métricas como la precisión, el MSE (Error Cuadrático Medio) y el AUC (Área Bajo la Curva) para comparar el rendimiento de los modelos 1D-CNN y 2D-CNN.
        - **Precisión:** Proporción de instancias correctamente predichas.
        - **MSE:** Media de los cuadrados de las diferencias entre los valores predichos y los reales.
        - **AUC:** Capacidad del modelo para diferenciar entre instancias defectuosas y no defectuosas a diferentes umbrales.
5. Quinta Fase: Explicabilidad con XAI
    - **Explicabilidad:** Utilizamos técnicas avanzadas de XAI como LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations) para proporcionar explicaciones transparentes y comprensibles

sobre el comportamiento de los modelos de predicción de defectos en el software.

- **LIME (Ribeiro, Singh & Guestrin, 2016)**: Proporciona explicaciones locales y comprensibles sobre la toma de decisiones de modelos complejos de ML, mejorando la interpretabilidad y la confianza en los modelos.
- **SHAP (Lundberg & Lee, 2017):** Proporciona una robusta interpretación de las contribuciones de las características en los modelos predictivos, facilitando una comprensión matizada del impacto de cada característica en la predicción de defectos.

La combinación de una arquitectura ligera de CNN con técnicas avanzadas de XAI mejora significativamente la interpretabilidad de los modelos de ML sin comprometer el rendimiento.

3.5 Comparación y Evaluación del Modelo LCNN en la Identificación de Defectos de Software.

El estudio incluye una comparación y evaluación del Modelo LCNN en la identificación de defectos de software, presentada en la Tabla 12.

Tabla 12. Comparación y Evaluación del Modelo LCNN en la Identificación de Defectos de Software: Mejoras en la Explicabilidad, Impacto, Propuestas y Resultados Relevantes"

Aspecto	Descripción
Mejoras en la Explicabilidad	Detalla la creación y uso de un modelo basado en XAI para el análisis de defectos de software, empleando LIME y SHAP para ofrecer interpretaciones más claras sobre la importancia de las características en los modelos de ML.
Marco LCNN	Lightweight CNN Architecture for Software Defect Feature Identification Using Explainable AI
Comparación con Enfoques Anteriores	El paper compara explícitamente la arquitectura LCNN, evaluando su desempeño en precisión y eficiencia en contraposición con los enfoques: Deep Representation and Ensemble Learning. **Contribución**: Solución al problema del desequilibrio de clases en SDP, evaluación en 12 datasets de NASA, **Limitación**: No aborda la predicción de defectos entre

	proyectos (cross-project). DAECNN-JDP. **Contribución**: Uso de autoencoder para eliminación de ruido, superioridad sobre 11 modelos de referencia en proyectos de código abierto. **Limitación**:Falta evaluación en proyectos comerciales, falta optimización de parámetros Transfer CNN Model:**Contribución**:Extracción de características semánticas transferibles, evaluación en 10 proyectos benchmark y 90 pares de tareas CPDP .
Impacto de las Mejoras	Se presentan resultados empíricos que demuestran cómo las técnicas de XAI mejoran la comprensión y precisión en la predicción de defectos de software. **Mejora en la Eficiencia de Identificación de Defectos**: El uso de un diseño de CNN ligero (LCNN) ha demostrado mejorar significativamente la eficiencia en la identificación de defectos de software. Esto se refleja en una capacidad mejorada para detectar y clasificar defectos de manera rápida y efectiva durante las fases de prueba. **Incremento en la Interpretabilidad con XAI**: La integración de técnicas de Inteligencia Artificial Explicable (XAI), como LIME y SHAP, ha aumentado la interpretabilidad del modelo LCNN. Estas herramientas proporcionan insights claros sobre la importancia de las características identificadas por el modelo, facilitando la comprensión del proceso de toma de decisiones del modelo en el contexto específico de la identificación de defectos de software. **Aplicabilidad Práctica Mejorada**: Al emplear el dataset PC1 Promise Repository y optimizar el modelo para ser compatible con CNNs, se ha mejorado la aplicabilidad práctica del modelo propuesto. Esto fortalece la relevancia del estudio al demostrar su capacidad para adaptarse y desempeñarse efectivamente en entornos de desarrollo de software reales. **Potencial Colaborativo con la Industria**: Existe un potencial significativo para que el modelo LCNN colabore con profesionales de la industria para abordar desafíos específicos relacionados con el desarrollo de software. Esto sugiere una vía para transferir y aplicar directamente las mejoras investigadas en contextos industriales y comerciales. **Limitaciones a superar**: Aunque se han logrado mejoras sustanciales, se reconoce la necesidad de validar la

	generalización del modelo a diferentes entornos de software y contextos específicos. Esto resalta áreas para futuras investigaciones y refinamientos del modelo.
Propuestas y Taxonomías	El documento describe el desarrollo de un modelo XAI para el análisis de defectos de software, utilizando LIME y SHAP para ofrecer interpretaciones claras de la importancia de las características en modelos de ML.
Resultados Relevantes	Se comparó el rendimiento del modelo LCNN con métodos tradicionales y otros modelos de última generación en términos de precisión, eficiencia y capacidad de interpretación. Los resultados destacan las ventajas significativas del enfoque propuesto. Se presentaron estudios de caso que ilustran la aplicabilidad del modelo LCNN en la identificación de defectos en diversos contextos de desarrollo de software, validando su utilidad práctica y su capacidad para mejorar la calidad del software.

4. Las Partes Prototípicas Profundas Facilitan la Identificación Morfológica de Cálculos Renales y son Competitivamente Robustas a las Perturbaciones Fotométricas (Flores-Araiza, D., et al. 2023).

La urolitiasis, una condición en la que se forman cálculos renales en el tracto urinario, representa un problema de salud pública en países desarrollados debido a su alta incidencia, afectando hasta un 10% de la población. La formación de estos cálculos está estrechamente vinculada a factores como la dieta, la edad, enfermedades crónicas y antecedentes familiares. El diagnóstico preciso de los tipos de cálculos renales mediante el análisis morfo-constitucional (MCA) es importante para determinar sus causas subyacentes y prescribir tratamientos personalizados que reduzcan las recurrencias.

4.1 Exploración de Partes Prototípicas en la Identificación de Cálculos Renales.

Sin embargo, el MCA tradicional es costoso, consume mucho tiempo y depende en gran medida de la pericia del especialista, lo que dificulta los diagnósticos durante intervenciones endoscópicas en tiempo real. Para abordar estas limitaciones, se han propuesto métodos basados en inteligencia artificial (AI), particularmente aprendizaje profundo (Deep Learning, DL), para automatizar y mejorar la clasificación de los cálculos renales. Aunque los modelos de DL han demostrado superar a los métodos no DL en precisión, su falta de explicabilidad es un obstáculo significativo en entornos clínicos.

4.2 Comparación de Métodos de Interpretabilidad en el Análisis de Imágenes de Cálculos Renales.

La propuesta utiliza partes prototípicas (PPs) aprendidas por el modelo de DL. Estas PPs no solo mejoran la precisión de la clasificación, sino que también proporcionan explicaciones visuales sobre las características morfológicas relevantes de cada imagen de cálculo renal. Esta capacidad de explicabilidad permite incrementar la aceptación de los especialistas médicos, quienes requieren comprender el razonamiento detrás de las recomendaciones de los sistemas de diagnóstico asistido por computadora (CADx).

4.3 Arquitectura del Modelo

La arquitectura del modelo implementado utiliza un extractor de características f(x)f(x)f(x), en este caso una CNN, para realizar la extracción de características semánticas de las imágenes de entrada xxx y aprender nnn prototipos por clase, que corresponden a partes en la imagen de entrada. Este primer paso de extracción de características implica la diversidad potencial y la calidad de las posibles partes prototípicas que se pueden aprender. Por lo tanto, se exploran tres arquitecturas de CNN (VGG16, ResNet50 y DenseNet201) para comparar los diferentes desempeños posibles obtenidos con nuestro enfoque y compararlos con CNNs puramente no interpretables presentes en el estado del arte para la identificación de cálculos renales.

Posteriormente, se añaden dos capas de convoluciones 1 × 1 para ajustar la profundidad de los mapas de activación de características a una profundidad seleccionada de 128 para la capa de PPs (prototipos), denominada capa g. La variable l se utiliza para indexar cada uno de los parches $128\times1\times1\times1$ $f(x)_l$ a través de las dimensiones espaciales. La capa de PPs g contiene nnn prototipos por clase, cada uno con dimensiones $128\times1\times1$.

Dado que un prototipo tiene el mismo número de canales pero una dimensión espacial menor que los mapas de características convolucionales f(x), podemos interpretar el prototipo como representando un patrón de activación prototípico de su clase y visualizarlo como un parche de la imagen de entrenamiento en la que aparece. La distancia entre p_j y cada uno de los parches espaciales 1×1 $f(x)_l$ del mapa de características convolucionales f(x) se mide mediante $d_{j,l} = \|p_j - f(x)_l\|_2^2$, estas distancias se convierten en similitudes $s_{j,l}$ mediante la fórmula:

$$s_{j,l} = log \frac{d_{j,l} + 1}{d_{j,l} + \epsilon}$$

donde ϵ es un valor pequeño para evitar la división por cero.

4.4 Métricas de Rendimiento Comparativo de Modelos ProtoPNet y Redes Neuronales Convolucionales (CNN) Base bajo Condiciones IID y OOD

La Tabla 8, presenta las métricas de rendimiento de los modelos implementados, así como de las redes neuronales convolucionales (CNN) utilizadas como "backbones" o arquitecturas base:

1. **Métricas de Rendimiento**: La tabla 8, muestra las métricas de rendimiento de los modelos implementados (ProtoPNet) y las CNN utilizadas como "backbones". Estas métricas podrían incluir precisión, precisión media, puntuación F1, entre otras.
2. **Tipos de Evaluaciones**:
    - **IID (Independent Identically Distributed)**: Estas pruebas se realizan con el mismo tipo de datos con los que el modelo fue entrenado. Los datos IID se refieren a datos que son independientes e idénticamente distribuidos, lo que significa que las muestras de datos son independientes unas de otras y provienen de la misma distribución de probabilidad.
    - **OOD (Out Of Distribution)**: Estas pruebas se realizan bajo perturbaciones aplicadas a los datos originales. Los datos OOD son aquellos que no siguen la misma distribución que los datos de entrenamiento, lo que representa un escenario donde el modelo debe generalizar a nuevas situaciones o datos no vistos.
3. **Modelos Pre-entrenados en ImageNet**: Todos los modelos comparados en la tabla 15 fueron pre-entrenados en ImageNet, una gran base de datos de imágenes comúnmente utilizada para entrenar modelos de visión por computadora. El pre-entrenamiento en ImageNet proporciona a los modelos una base sólida de características visuales antes de ser ajustados para tareas específicas, como la clasificación de cálculos renales en este caso.

Tabla 15. Métricas de rendimiento.(Flores-Araiza et al. (2024)).

Model		IID Accuracy (%)	IID Precision (%)	IID F1 (%)	OOD Accuracy (%)	OOD Precision (%)	OOD F1 (%)
DenseNet201		**89.67±3.60**	**90.51±3.60**	**89.47±3.46**	**58.44±28.06**	60.22±29.62	**54.54±32.73**
ResNet50		83.40±5.32	85.58±4.30	83.11±5.21	45.12±20.82	49.16±24.08	38.25±23.47
Vgg16		81.47±2.72	84.71±2.64	80.62±3.84	58.35±22.62	**63.73±19.93**	54.38±26.30
CNNs Average		84.84±5.29	86.93±4.35	84.40±5.60	53.97±24.62	57.70±25.40	49.06±28.56
DenseNet201	1 PPs	86.29±1.91	87.54±1.34	86.25±1.78	60.31±20.32	70.29±14.38	57.45±23.23
	3 PPs	85.19±1.50	86.08±1.27	85.21±1.42	59.97±19.91	69.21±12.59	57.66±22.14
	10 PPs	87.29±0.92	87.99±0.94	87.23±0.92	**61.20±19.33**	**71.26±11.92**	**59.01±21.37**
ResNet50	1 PPs	**88.21±2.07**	**88.61±1.88**	**88.21±2.05**	58.26±24.23	66.48±18.53	55.16±27.50
	3 PPs	86.66±1.37	87.11±1.55	86.62±1.45	57.90±21.82	63.24±19.16	54.21±25.32
	10 PPs	85.44±1.44	86.16±1.15	85.43±1.42	56.31±22.66	64.81±16.55	53.77±25.52
Vgg16	1 PPs	81.78±1.60	83.38±1.94	81.82±1.62	51.78±19.02	53.82±19.40	46.63±22.77
	3 PPs	82.21±3.33	82.92±3.82	81.84±3.75	51.23±19.40	54.09±19.80	45.88±23.20
	10 PPs	82.08±0.90	83.39±0.92	82.13±0.87	54.69±18.05	56.48±17.56	49.89±21.51
ProtoPNets Average		85.02±2.83	85.91±2.66	84.97±2.89	56.85±20.63	63.30±17.90	53.30±23.83

Los modelos ProtoPNet (Prototype Network) son una clase de redes neuronales diseñadas para ser interpretables. A diferencia de las redes neuronales convolucionales (CNN) tradicionales, que a menudo son consideradas "cajas negras" debido a su falta de interpretabilidad, los modelos ProtoPNet permiten a los usuarios entender y visualizar cómo y por qué se realizan las decisiones de clasificación.

La arquitectura de un modelo ProtoPNet se basa en los siguientes componentes clave:

- **Extracción de características**: Se utiliza una CNN para extraer características semánticas de las imágenes de entrada. Esta CNN actúa como un extractor de características que transforma las imágenes en mapas de activación de características.
- **Prototipos por clase**: El modelo aprende prototipos (representaciones prototípicas) para cada clase. Estos prototipos son pequeñas partes de las imágenes de entrada que representan patrones característicos de cada clase.
- **Cálculo de similitud**: Para clasificar una nueva imagen, el modelo compara sus características extraídas con los prototipos aprendidos, calculando una puntuación de similitud. La similitud entre las características de la imagen y los prototipos se mide usando una métrica de distancia.
- **Mapas de similitud y saliencia**: Las puntuaciones de similitud se organizan en mapas de similitud que se pueden superponer a la imagen de entrada para generar mapas de saliencia. Estos mapas visualizan qué partes de la imagen son más importantes para la clasificación.
- **Clasificación final**: Las similitudes calculadas se combinan en una capa completamente conectada, ponderando las similitudes para obtener una salida por clase. Esta salida se normaliza usando una función softmax para obtener las probabilidades de clasificación.

El rendimiento de los modelos ProtoPNet implementados es comparable con sus correspondientes modelos de "backbone" CNN no interpretables, como se aprecia en la Tabla 8. La precisión promedio de las implementaciones se mantiene competitiva, con una diferencia no mayor a ≤ 0.18% en comparación con la precisión promedio de los modelos CNN de referencia. Los valores medios y las desviaciones estándar mostradas en la Tabla 15 se calculan en base a las configuraciones de los modelos descritas en la misma.

La Tabla 15 muestra que los modelos interpretables superan a las CNN tradicionales del estado del arte en la clasificación de cálculos renales cuando se evalúan en imágenes de entrada perturbadas. Específicamente, lograron un 1.9% más de precisión, un 5.7% más de precisión y un 3.2% más de puntuación F1 en comparación con los modelos tradicionales.

4.5 Avances más recientes en la mejora de la explicabilidad de modelos 'caja negra'

Este estudio explora cómo los modelos basados en PPs, en particular la arquitectura ProtoPNet, representan un avance significativo al proporcionar taxonomías visuales detalladas y métricas de desempeño comparativo con modelos CNN tradicionales como se puede ver en la tabla 16. Estos desarrollos no solo permiten mejorar la precisión y robustez en la identificación de cálculos renales, sino que también establecen un estándar para la evaluación de modelos interpretables en aplicaciones médicas críticas.

En respuesta a la pregunta de investigación presentada en el punto 3.2, se derivan los siguientes aspectos:

1. Identificación de Nuevas Metodologías:
    - El estudio introduce el concepto de Prototypical Parts (PPs), que son partes prototípicas aprendidas por un modelo de Deep Learning (DL). Estas PPs no solo mejoran la precisión de la clasificación de cálculos renales, sino que también proporcionan explicaciones visuales sobre las características morfológicas relevantes de cada imagen. Esto permite una mayor interpretabilidad del modelo en comparación con enfoques anteriores que carecen de esta capacidad explícita.
2. Comparación con enfoques anteriores:
    - **Comparación con Métodos Tradicionales:**
        - En contraste con métodos tradicionales de Aprendizaje Profundo que son considerados 'caja negra' debido a su falta de explicabilidad, los modelos basados en PPs permiten una interpretación más clara de las decisiones del modelo. Esto se ilustra mediante la comparación directa con modelos convencionales que muestran resultados competitivos en

términos de precisión, pero sin ofrecer explicaciones transparentes sobre el proceso de toma de decisiones.

3. Impacto de las mejoras:
    - **Propuestas y Taxonomías:**
        - El estudio propone una nueva arquitectura de red neuronal, ProtoPNet, que implementa PPs para mejorar la explicabilidad en la clasificación de cálculos renales. Esta propuesta se posiciona como un avance significativo al proporcionar taxonomías visuales detalladas de las PPs aprendidas, lo cual es crucial para entender cómo y por qué se realizan las clasificaciones.
        - Además, el artículo introduce un marco de evaluación que incluye métricas cuantitativas de desempeño comparativo con modelos CNN tradicionales, estableciendo un estándar para la evaluación de modelos interpretables en aplicaciones médicas.
    - **Resultados Relevantes:**
        - Se presentan casos de uso práctico donde los modelos ProtoPNet muestran una mejora significativa en la precisión y robustez frente a perturbaciones fotométricas en comparación con los modelos CNN no interpretables del estado del arte. Esto demuestra el impacto directo de la explicabilidad en la capacidad de los modelos para generalizar datos no vistos y mejorar la confianza de los profesionales médicos en las recomendaciones del sistema CADx.

Tabla 16. Comparación de Avances en la Explicabilidad de Modelos 'Caja Negra' en la Identificación de Cálculos Renales

Aspecto	Descripción
Mejoras en la Explicabilidad	Introducción de Prototypical Parts (PPs) para mejorar la explicabilidad de modelos 'caja negra' en la identificación de cálculos renales.
Marco PPs	Uso de PPs para proporcionar explicaciones visuales detalladas y transparentes sobre las decisiones de clasificación de modelos de Deep Learning.
Comparación con Enfoques Anteriores	Comparación directa con métodos tradicionales de Aprendizaje Profundo que carecen de explicabilidad, demostrando

	avances significativos en transparencia y entendimiento.
Impacto de las Mejoras	Aumento en la aceptación y confianza por parte de especialistas médicos al entender el razonamiento detrás de las recomendaciones de sistemas de diagnóstico asistido por computadora (CADx).
Propuestas y Taxonomías	Desarrollo de un nuevo marco de evaluación basado en PPs y taxonomías visuales que facilitan la adopción y aplicación de modelos interpretables en diagnóstico médico.
Resultados Relevantes	Mejoras demostradas en precisión y robustez frente a perturbaciones, validando la utilidad clínica de los modelos interpretables en diagnóstico de cálculos renales.

5. RoCourseNet: Entrenamiento Robusto de un Modelo de Recurso Consciente de Predicciones (Guo et al., 2023).

5.1 Arquitectura del modelo

RoCourseNet es un marco de entrenamiento de extremo a extremo diseñado para generar simultáneamente predicciones precisas y recursos correspondientes (explicaciones contrafactuales o CF) que sean robustos frente a cambios en el modelo inducidos por desplazamientos en el conjunto de datos de entrenamiento. Este marco se describe en dos etapas principales: el problema del atacante y el problema del defensor.

La **Fig.8** ilustra el proceso de generación de recursos robustos. **Fig.8 (a)** Dado un punto de datos de entrada x, los métodos de explicación CF generan un nuevo recurso $x$cf que se encuentra en el lado opuesto de la frontera de decisión $f(\cdot; \theta)$. **Fig.8 (b)** A medida que se disponen de nuevos datos, la frontera de decisión del modelo de aprendizaje automático se actualiza como $f(\cdot; \theta')$. Esta nueva frontera desplazada $f(\cdot; \theta')$ invalida el recurso elegido $x$cf (ya que x y $x$cf quedan en el mismo lado del modelo desplazado $f(\cdot; \theta')$). **Fig.8 (c)** Sin embargo, los métodos robustos de explicación CF generan un recurso robusto $\hat{x}$cf para el dato de entrada x anticipando el modelo desplazado futuro $f(\cdot; \theta')$.

Figura 8.Proceso de generación de recurso robusto en métodos de explicación CF (Guo et al. (2023).

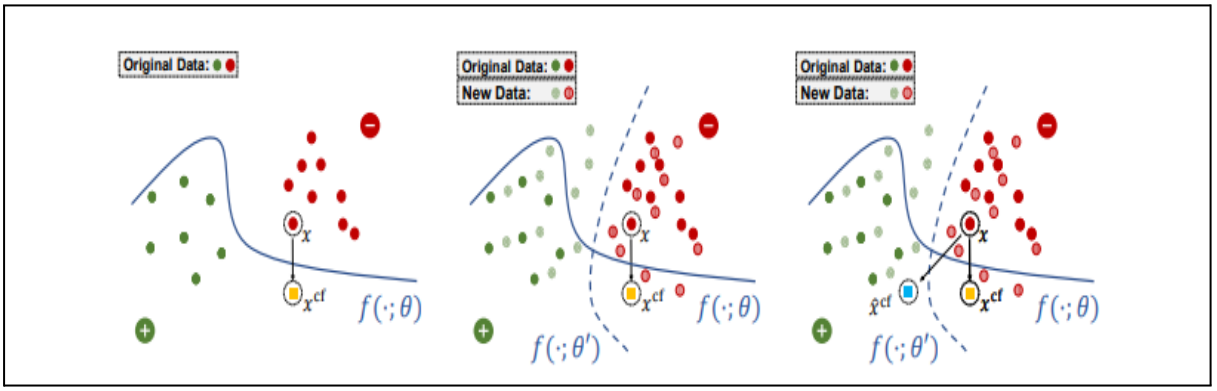
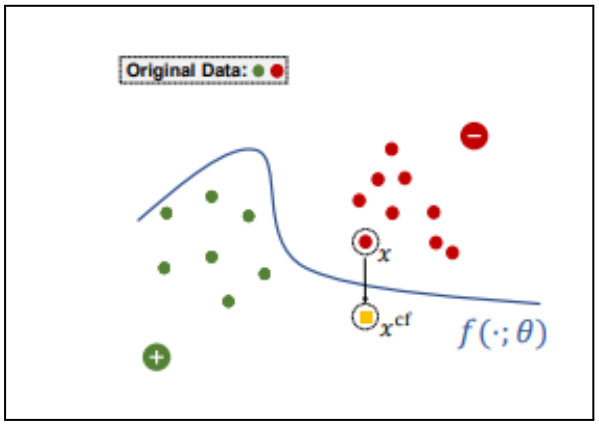
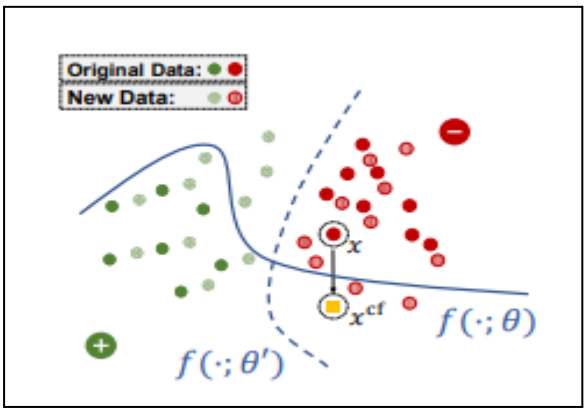


Figura 8 (a) Generación de Recurso CF (Guo et al. (2023).



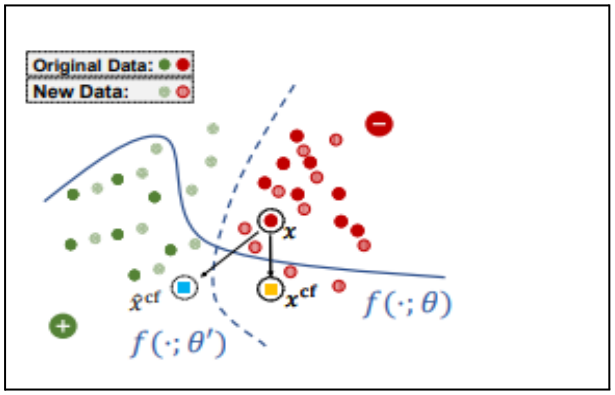
(a) La frontera de decisión del modelo $f(\cdot, \theta)$ entrenado con los datos originales, y un recurso x cf del ejemplo x.

Figura 8(b)Actualización de la Frontera de Decisión (Guo et al. (2023).



(b) Frontera de decisión actualizada del nuevo modelo reentrenado $f(\cdot, \theta')$ con datos recién disponibles (o una distribución de datos modificada).

Figura 8 (c) Generación de Recurso Robusto(Guo et al. (2023) .



(c) Bajo los cambios en la distribución de datos y el modelo, el recurso *x* cf se vuelve inválido, pero *x*ˆ cf es válido. Llamamos a *x*ˆ cf un recurso robusto.

Etapa 1: Problema del Atacante

Se aborda como un problema de optimización en dos niveles. Utiliza el algoritmo Virtual Data Shift (VDS) basado en gradientes para simular los peores cambios posibles en los datos de entrenamiento y evaluar la robustez de las explicaciones CF.

- **Problema del Atacante en Dos Niveles:** Se propone un problema de optimización en dos niveles para encontrar el peor cambio en los datos de entrenamiento que genere un modelo adversarialmente alterado, reduciendo la validez de las explicaciones CF.

$$\theta'_{adv} = \arg\max_{\theta' \in F} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[L\left(f(x_{cf_i}; \theta'), 1 - f(x_i; \theta) \right) \right]$$

Donde:

- $f(x, \theta)$ es la predicción del modelo en el punto x con parámetros θ.
- x_{cf} representa una explicación contrafactual para el punto de entrada x.
- $F = \{\theta' | \theta + \delta\theta\}$ es un conjunto plausible de parámetros de modelos desplazados.

- **Virtual Data Shift (VDS):**

  Para resolver el problema del atacante, se propone el algoritmo de Virtual Data Shift (VDS), un enfoque basado en gradientes que simula el desplazamiento de datos adversos:

$$\delta^* = \arg\max_{\delta, \forall \delta_i \in \Delta} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[L\left(f(x_{cf_i}; \theta_{opt}(\delta)), 1 - f(x_i; \theta) \right) \right]$$

Donde:

- $\Delta = \{\delta \in \mathbb{R}^n | \|\delta\|_\infty \leq \epsilon\}$ es una bola ℓ^∞ que representa el espacio de desplazamiento de datos.
- $\theta_{opt}(\delta) = \arg\min_{\theta'} \frac{1}{N} \sum_{(x_i, y_i) \in D} L\left(f(x_i + \delta_i; \theta'), y_i \right)$.

Un algoritmo diseñado para resolver este problema de dos niveles mediante la creación de desplazamientos de datos virtuales, que imitan los peores cambios posibles en el conjunto de datos.

Etapa 2: Problema del Defensor

Expande el enfoque del atacante a un problema de optimización en tres niveles. El objetivo es optimizar simultáneamente la precisión de las predicciones y la robustez de las explicaciones CF contra cambios adversos en los datos de entrenamiento. RoCourseNet integra el método VDS en su marco de entrenamiento para mejorar la estabilidad de las explicaciones CF.

- **Problema del Defensor en Tres Niveles:** Se plantea un problema de optimización en tres niveles basado en el problema de dos niveles del atacante, con el objetivo de generar predicciones precisas y explicaciones CF robustas simultáneamente.

$$\theta_{opt} = \arg\min_{\theta'} \frac{1}{N} \sum_{(x_i, y_i) \in D_{shifted}} L\left(f(x_i; \theta'), y_i \right)$$

Donde $D_{shifted}$ es el conjunto de datos desplazado derivado del desplazamiento adverso.

- **RoCourseNet Training Framework:** Un marco de entrenamiento que optimiza este problema de tres niveles, integrando el método VDS para mejorar la robustez de las explicaciones CF.

$$\min_{\theta,\theta_g} [\lambda_1 \cdot L1 + \max_{\delta,\forall\delta_i\in\Delta} (\lambda_2 \cdot L2, 1 - f(x_i;\theta))]$$

Donde:

- $L1$ es la pérdida de predicción.
- $L2$ es la pérdida de validez robusta de las CF generadas.
- θ_g es el generador de recursos CF.
- $\theta_{opt}(\delta)$ es el modelo ajustado por el desplazamiento de datos adversos.

5.2 Marco de entrenamiento RoCourseNet

Elección de la técnica de explicación CF: RoCourseNet requiere una técnica de explicación CF adecuada que pueda generar recursos para los puntos de datos de entrada. Dado que los métodos de explicación CF post hoc no son adecuados para este marco debido a su paradigma de funcionamiento y a las regulaciones de derecho a la explicación como GDPR, se elige CounterNet.
CounterNet se diferencia de los enfoques post hoc al entrenar conjuntamente predicciones y recursos, mejorando el equilibrio entre el costo y la invalidez en comparación con los métodos post hoc.

Función objetivo de RoCourseNet: La función objetivo de RoCourseNet se formula como un problema de minimización-maximización-minimización.

Tiene tres objetivos principales: (i) alta precisión predictiva, (ii) alta validez robusta de las explicaciones CF frente a modelos adversarialmente alterados, y (iii) baja proximidad, es decir, cambios mínimos necesarios para modificar una instancia de entrada a su recurso correspondiente.

RoCourseNet optimiza estos objetivos simultáneamente ajustando los parámetros de su predictor $f(.;\theta)$ y generador de recursos $g(.;\theta_g)$ utilizando un enfoque de optimización en tres niveles.

$$\underset{\theta,\theta_g}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} \left[\lambda_1 \cdot \underbrace{\mathcal{L}\left(f(x_i;\theta), y_i\right)}_{\text{Prediction Loss } (L_1)} + \lambda_3 \cdot \underbrace{\mathcal{L}\left(x_i, x_i^{\text{cf}}\right)}_{\text{Proximity Loss } (L_3)} \right] \right.$$

$$\left. + \max_{\boldsymbol{\delta},\forall\delta_i\in\Delta} \frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} \left[\lambda_2 \cdot \underbrace{\mathcal{L}\left(f\left(x_i^{\text{cf}}; \theta'_{opt}(\boldsymbol{\delta})\right), 1 - f\left(x_i;\theta\right)\right)}_{\text{Robust Validity Loss } (L_2)} \right] \right)$$

$$s.t., \theta'_{opt}(\boldsymbol{\delta}) = \underset{\theta'}{\operatorname{argmin}} \frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} \left[\mathcal{L}\left(f(x_i+\delta_i;\theta'), y_i\right) \right],$$

$$x_i^{\text{cf}} = g(x_i;\theta_g).$$

Entrenamiento de RoCourseNet: Para resolver la formulación min-max, RoCourseNet utiliza un enfoque de descenso de coordenadas por bloques y un algoritmo denominado Virtual Data Shift (VDS) para manejar los desplazamientos de datos adversos. El VDS simula los peores cambios posibles en los datos de entrenamiento, mientras que el descenso de coordenadas por bloques optimiza la función objetivo en dos etapas: primero, para la precisión predictiva y luego, para la calidad de las explicaciones CF.

RoCourseNet se adapta dinámicamente la intensidad del adversario aumentando linealmente las restricciones de perturbación a lo largo del entrenamiento. Este enfoque se basa en el aprendizaje adversarial curricular, mejorando la convergencia del entrenamiento.

5.3 Comparativa Experimental de RoCourseNet con Baselines de Generación de Recursos Robustos en Modelos Predictivos

RoCourseNet representa un avance significativo al ser el primer método que optimiza un modelo end-to-end para la generación de predicciones y recursos robustos. Dado que no existen enfoques previos directamente comparables, se ha comparado RoCourseNet con cuatro baselines de estado del arte:

1. **VanillaCF** (Wachter, Mittelstadt, & Russell, 2017): Este método post hoc no paramétrico se aplica después del entrenamiento del modelo de aprendizaje automático. Se enfoca en optimizar la validez y proximidad de las explicaciones contrafactuales, asegurando que sean válidas y cercanas al ejemplo original en términos de características y distribución de datos.
2. **ROAR-LIME** (Upadhyay, Joshi, & Lakkaraju, 2021): Genera recursos robustos al perturbar los parámetros de modelos lineales locales. Utiliza modelos lineales

aproximados para modelar cómo las pequeñas variaciones en los parámetros afectan las predicciones, sin necesidad de conocer el modelo original en su totalidad.

3. **RBR (Robust Bias Regularization)** (Nguyen et al., 2022): Este método se centra en regularizar el sesgo de los recursos generados para garantizar su validez y utilidad en diferentes contextos y modelos. Controla la generación e interpretación de recursos adicionales en relación con las predicciones del modelo.
4. **CounterNet** (Guo, Nguyen, & Yadav, 2021): Es un modelo end-to-end que simultáneamente genera predicciones y explicaciones contrafactuales durante el entrenamiento. A diferencia de los métodos post hoc, CounterNet optimiza la calidad de las explicaciones dentro del proceso de entrenamiento del modelo principal, mejorando la coherencia y especificidad de las predicciones del modelo.

RoCourseNet se evaluó exhaustivamente en tres conjuntos de datos del mundo real:

- **Loan**: Registros de solicitudes de préstamos que abarcan varios años en los EE. UU., con aproximadamente 450,000 puntos de datos.
- **German Credit**: Datos que capturan cambios en corrección, evaluando el puntaje crediticio de clientes.
- **Student**: Registros de estudiantes que capturan cambios geoespaciales en el entorno escolar.

El procedimiento de evaluación dividió cada conjunto de datos en subconjuntos originales y sus respectivos desplazamientos, permitiendo la medición de la robustez de los recursos generados frente a estos cambios. Cada subconjunto original fue utilizado para entrenar un modelo separado, cuya robustez fue evaluada en los conjuntos de prueba utilizando métricas como la validez, la validez robusta y la proximidad.

5.4 Explicabilidad de modelos de Caja Negra con RoCourseNet.

En respuesta a la pregunta de investigación planteada en el punto 3.2, este análisis examina los avances recientes en la explicabilidad de modelos 'caja negra', centrándose específicamente en el método RoCourseNet. Este marco de entrenamiento end-to-end ha sido diseñado para mejorar significativamente la generación de explicaciones contrafactuales robustas en modelos predictivos. Utilizando un enfoque tri-level de optimización y técnicas innovadoras de entrenamiento adversarial, RoCourseNet representa una metodología avanzada que integra de manera coherente y efectiva la generación de predicciones con la producción de explicaciones. Finalmente, el análisis concluye con la Tabla 18, que proporciona una comparación detallada de RoCourseNet con métodos tradicionales y discute su impacto en la explicabilidad de modelos 'caja negra'.

Avances recientes en la explicabilidad de modelos 'Caja Negra'

RoCourseNet representa un avance significativo al introducir un marco de entrenamiento end-to-end diseñado específicamente para mejorar la generación de explicaciones contrafactuales robustas en modelos predictivos. Utiliza un enfoque tri-level de optimización y técnicas novedosas de entrenamiento adversarial para abordar este desafío. Este enfoque innovador se destaca como una nueva metodología en el campo de la explicabilidad de modelos 'caja negra', ya que integra la generación de predicciones con la producción de explicaciones de manera coherente y eficaz.

Comparación con enfoques anteriores

En su análisis comparativo, RoCourseNet contrasta con métodos tradicionales como VanillaCF, ROAR-LIME y RBR. A diferencia de estos enfoques que son post-hoc o utilizan métodos específicos como modelos lineales locales o núcleos gaussianos, RoCourseNet se posiciona como el primer método end-to-end para generar explicaciones robustas. Esta comparación explícita subraya las mejoras significativas de RoCourseNet en términos de validez robusta y balance en el trade-off entre costo y validez, demostrando su superioridad frente a los métodos existentes.

Impacto de las mejoras

RoCourseNet no solo propone un marco teórico avanzado, sino que también ofrece resultados prácticos tangibles. Se han desarrollado propuestas innovadoras y un framework robusto que puede ser generalizado para aplicarse con cualquier método paramétrico de explicación contrafactual. Este enfoque tiene el potencial de transformar la forma en que se entienden y se utilizan las explicaciones de modelos predictivos en diversas aplicaciones prácticas, mejorando la transparencia y la confiabilidad de los sistemas de inteligencia artificial.

Resultados relevantes

Los resultados empíricos de RoCourseNet muestran mejoras significativas en términos de la calidad y la robustez de las explicaciones contrafactuales generadas, validando su eficacia en escenarios del mundo real como préstamos, evaluación crediticia y predicción del rendimiento estudiantil. Estos estudios de caso proporcionan evidencia concreta del impacto práctico de las mejoras introducidas por RoCourseNet en la explicabilidad de los modelos 'caja negra'.

Tabla 18. Comparación de RoCourseNet con Métodos Tradicionales y su Impacto en la Explicabilidad de Modelos 'Caja Negra'.

Aspecto	Descripción
Mejoras en la Explicabilidad	RoCourseNet introduce un marco end-to-end que optimiza la generación de predicciones y explicaciones contrafactuales robustas en modelos de machine learning 'caja negra'.
Marco end-to-end	Utiliza un enfoque tri-level de optimización y técnicas de entrenamiento adversarial para mejorar la transparencia y la robustez de las explicaciones generadas.
Comparación con Enfoques Anteriores	Contrasta con métodos tradicionales como VanillaCF, ROAR-LIME y RBR, demostrando mejoras significativas en validez robusta y en el trade-off entre costo y validez.
Impacto de las Mejoras	Proporciona un framework generalizable aplicable a métodos paramétricos de explicación contrafactual, elevando la transparencia y fiabilidad de los sistemas de inteligencia artificial.
Propuestas y Taxonomías	Introduce propuestas innovadoras y frameworks robustos que han transformado la interpretación y la confiabilidad en aplicaciones prácticas como evaluación crediticia y educación.
Resultados Relevantes	Demuestra mejoras significativas en la calidad y robustez de las explicaciones contrafactuales, validando su eficacia en escenarios del mundo real y aplicaciones prácticas diversas.

6. Explicabilidad Federada para la Caracterización de Anomalías en Redes (Sáez-de-Cámara, Flores, Arellano, Urbieta, & Zurutuza, 2023).

A continuación se revisan trabajos recientes que abordan la explicabilidad en ciberseguridad, tanto en entornos de federated learning (FL) como en entornos no federados, destacando las

metodologías y enfoques utilizados para interpretar las predicciones de los modelos y mejorar la confianza de los usuarios en estas tecnologías avanzadas

6.1 Explicabilidad para la Detección de Anomalías o Ataques en Ciberseguridad (Configuraciones no Federadas)

La mayoría de los trabajos sobre técnicas de Inteligencia Artificial Explicable (XAI) en ciberseguridad se centran en la visualización y verificación de modelos o predicciones. A continuación, se presentan algunos estudios relevantes:

Wang et al. (2020): Utilizan SHAP para proporcionar explicaciones locales y globales de los Sistemas de Detección de Intrusiones (IDS) y ayudar a los analistas de seguridad a interpretar las predicciones. Evaluaron dos modelos supervisados en el conjunto de datos NSL-KDD [1], mostrando diferentes patrones de valores SHAP según el tipo de ataque. Sin embargo, se limitaron a la visualización sin profundizar en análisis adicionales.

Antwarg et al. (2020): Utilizan SHAP para explicar anomalías detectadas por un modelo de autoencoder no supervisado, identificando características con altos errores de reconstrucción. Evaluaron su método en el conjunto de datos KDD Cup 1999, entre otros, y visualizaron las explicaciones para facilitar la comprensión y clasificación de las anomalías.

Liu et al. (2021): Presentan FAIXID, un marco para añadir explicabilidad a IDS en varias capas: limpieza de datos, explicación de modelos supervisados entrenados, explicaciones locales de predicciones y presentación de resultados a analistas de seguridad a través de diversas visualizaciones adaptadas a sus roles.

Rao et al. (2021): Entrenan un bosque de aislamiento en el conjunto de datos NSL-KDD para clasificar muestras normales y anómalas. Utilizan SHAP y LIME para extraer y visualizar explicaciones, generando automáticamente etiquetas para los ataques asignando el nombre de la característica más importante en la predicción.

Otros estudios combinan explicabilidad con análisis adicionales para extraer más información de las anomalías detectadas o clases predichas:

Nguyen et al. (2020): Presentan GEE, un autoencoder variacional explicable para la detección de anomalías en redes, evaluado con datos NetFlow del conjunto UGR16. Utilizan una técnica basada en gradientes para explicar las anomalías y agruparlas por similitud,

[1] NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB

aunque este punto no se explora en profundidad y el método de gradientes es específico para el modelo VAE.

Liyanage et al. (2021): Desarrollan un marco para caracterizar ataques a partir de anomalías en el flujo de red utilizando minería de conjuntos frecuentes (FIM), en lugar de técnicas XAI o explicaciones basadas en gradientes de GEE. Algunos pasos de la minería requieren datos etiquetados.

Barnard et al. (2021): Proponen un IDS de red en dos etapas: primero, entrenan un modelo XGBoost supervisado para la clasificación binaria de datos de flujo de red y usan SHAP para explicar las predicciones. Luego, un autoencoder usa las explicaciones de SHAP como entrada para distinguir comportamientos normales de los desconocidos. Evaluaron su propuesta en el conjunto de datos NSL-KDD.

Sudheera et al. (2021): Desarrollan ADEPT, un marco para la detección de anomalías en flujos de red y la identificación de etapas de ataque en una red IoT distribuida. Funciona en tres fases: detección local de anomalías por cada cliente, envío de flujos anómalos al servidor central, y minería de datos con FIM en el servidor central para clasificar los flujos maliciosos en etapas de ataque. Aunque no consideran la explicabilidad, los patrones extraídos son interpretables. A diferencia de las arquitecturas de FL, su enfoque envía datos sensibles al servidor central, mientras que FL podría ofrecer mayor privacidad y eficiencia en la comunicación.

6.2 Explicabilidad en Ciberseguridad en Entornos de Aprendizaje Federado

Haffar et al. (2022) utilizan bosques aleatorios (RF) como sustitutos del modelo supervisado de aprendizaje federado (FL). Cada cliente en la red entrena un RF con sus datos locales. Cuando el modelo de FL clasifica erróneamente una muestra, se utilizan los árboles en el RF para calcular la importancia de las características, detectando y explicando ataques contra el proceso de entrenamiento del modelo de FL. Las explicaciones se realizan a nivel de cliente y requieren datos de entrenamiento etiquetados. Cada cliente tiene su propio modelo explicativo, lo que puede dificultar la interpretación global de las explicaciones, ya que no están entrenados de manera federada. Su enfoque se centra en detectar ataques contra el proceso de entrenamiento de FL, no en explicar y caracterizar las predicciones.

Huong et al. (2022) proponen una arquitectura de detección de anomalías basada en FL para sistemas de control industrial. Utilizan SHAP para interpretar y verificar el modelo de FL entrenado, proporcionando visualizaciones como herramienta de apoyo para expertos en el dominio. El modelo explicativo de SHAP no se entrena de manera federada. SHAP necesita muestras de datos de fondo como referencia; sin embargo, los autores no explican cómo se

extrae esta referencia, lo cual es crucial debido a la naturaleza distribuida de los datos en entornos de FL.

La mayoría de la literatura sobre técnicas de inteligencia artificial explicable (XAI) se enfoca en la visualización y verificación de modelos. Los trabajos que utilizan explicaciones para funcionalidades adicionales, como agrupar o caracterizar anomalías, son raros y están diseñados para arquitecturas centralizadas o distribuidas, no para el aprendizaje federado (FL). Además, estos trabajos suelen necesitar datos etiquetados, lo cual no siempre es práctico.

En FL, algunas investigaciones utilizan XAI para verificar el proceso de entrenamiento o visualizar y verificar el modelo, pero no entrenan los modelos explicativos de manera federada, lo que complica la interpretación de las explicaciones a través de la red federada. Ninguno de los trabajos considera cómo seleccionar una línea de base común para SHAP en FL, lo cual es importante para generar explicaciones consistentes entre los clientes federados.

6.3 Arquitectura del modelo propuesto

Los componentes del método propuesto, están divididos en tres bloques principales. El enfoque principal se centra en el tercer bloque, que aborda el entrenamiento del modelo explicador federado y la caracterización de anomalías.

El último bloque incluye dos etapas realizadas de manera federada: el entrenamiento del modelo explicador y la caracterización de las anomalías. Para entrenar el modelo explicador utilizando Kernel SHAP, se requieren dos entradas principales: el modelo de detección de anomalías global f entrenado mediante FL, común a todos los clientes, y un conjunto de datos de fondo representativo. Dado que los datos en entornos FL están distribuidos entre todos los clientes y no se comparten centralmente, se emplea una versión adaptada de k-means federado (k-FED) (Dennis, Li, & Smith, 2021) para generar un conjunto de datos de fondo común a partir del conjunto de datos distribuido.

El proceso de caracterización de anomalías es la segunda etapa crítica que también se realiza de manera federada. Utiliza las explicaciones generadas por SHAP para las muestras anómalas, ϕi, que muestran la importancia de cada característica. Además, se emplean los datos procesados y sin procesar de las muestras anómalas, incluyendo características como direcciones IP y marcas de tiempo, que no se usan para entrenar modelos de ML para evitar correlaciones espurias, pero que son importantes para los analistas de seguridad.

Mediante el uso de FL, se asegura que todos los clientes puedan identificar y conocer las diferentes actividades anómalas en toda la red federada, incluso si cada cliente ha estado expuesto a diferentes tipos de ataques. Específicamente, k-FED (Dennis, Li, & Smith, 2021)

se utiliza nuevamente para agrupar los resultados de explicabilidad en cada cliente y compartirlos con otros pares en la red, garantizando que todos puedan hacer referencia a las mismas etiquetas de agrupación para las instancias anómalas detectadas en la red federada.

6.4 Proceso de entrenamiento

Entrenamiento del explainer SHAP federado:

- Se emplea el modelo explicador Kernel SHAP (SHAP por kernel) para explicar las decisiones de un modelo de detección de anomalías entrenado de forma federada.
- El proceso de entrenamiento requiere un conjunto de clientes Z, donde cada cliente tiene su propio número local de clusters $k(z)$ y hay un número global de clusters k.
- Debido a la complejidad computacional, se eligen valores pequeños para k en relación con el tamaño del conjunto de datos de entrenamiento. Se exploran dos valores: $k = 5$ y $k = 20$.
- Cada cliente calcula valores SHAP sobre muestras anómalas identificadas, normalizándolos posteriormente.

Visualización y análisis de SHAP values:

- Se realiza una visualización bidimensional de los valores SHAP generados centralmente utilizando la técnica de reducción de dimensionalidad UMAP.
- Los resultados muestran diferencias en la distribución de los clusters para $k = 5$ y $k = 20$, destacando una mayor definición de clusters para $k = 20$.
- Cada punto anómalo se colorea según una etiqueta de ataque, aunque estas etiquetas se usan sólo con propósitos de visualización y no para el entrenamiento del modelo.

Aplicación en conjunto de datos de paquete y flujo:

- Para el conjunto de datos a nivel de paquete, se utilizan 11 clientes en el entrenamiento del modelo federado, con 2 de ellos recibiendo ataques como el tráfico C&C de Mirai y escaneos de Nmap.
- En el conjunto de datos a nivel de flujo (N-BaIoT), se simulan 15 clientes donde cada tipo de ataque se asigna a un cliente simulado. Se comparan muestras benignas para calcular los valores SHAP de las anomalías.

Clustering de anomalías federado:

- Se realiza la agrupación de anomalías federada utilizando los valores SHAP obtenidos con $k = 20$ como referencia para ambos conjuntos de datos.

- Se estima el número de clusters anómalos locales $k(z)$ utilizando HDBSCAN y se computa k' como el máximo de todos los $k(z)$ recibidos.
- Se realizan múltiples pruebas de k-means federado para determinar el número óptimo de clusters, evaluando el índice CH para seleccionar el mejor ajuste.

Validación y resultados:

- Se muestran métricas de validación de agrupamiento como el índice CH, destacando la calidad del agrupamiento federado en comparación con un enfoque centralizado.
- Se elige el número óptimo de clusters basado en métricas de validación interna no supervisadas, evitando el uso de resultados de agrupamiento de referencia o la centralización de datos.

6.5 Análisis del Método Propuesto

El método propuesto utiliza Kernel SHAP en un entorno federado para mejorar la explicabilidad de modelos de detección de anomalías en redes, como se detalla en la Tabla 19. Este enfoque se destaca por su innovación al aplicar SHAP en contextos distribuidos y heterogéneos, permitiendo una interpretación más robusta de las predicciones del modelo (véase Tabla 20, Mejoras en la Explicabilidad y Marco k-means federado).

El análisis comparativo con enfoques anteriores subraya cómo el enfoque federado supera desafíos críticos como la centralización de datos y la escalabilidad, aprovechando la distribución de datos entre múltiples clientes (véase Tabla 20, Comparación con Enfoques Anteriores). Esta característica no solo mejora la escalabilidad al distribuir la carga de trabajo, sino que también fortalece la privacidad al mantener los datos sensibles en los dispositivos locales (véase Tabla 20, Impacto de las Mejoras).

6.6 Identificación de Nuevas Metodologías

El método propuesto utiliza Kernel SHAP en un entorno federado para mejorar la explicabilidad de modelos de detección de anomalías en redes. Esto constituye una metodología innovadora en la aplicación de SHAP en un contexto distribuido y federado, donde los modelos pueden considerarse 'caja negra' debido a la heterogeneidad y distribución de los datos.

6.7 Comparación con Enfoques Anteriores

El texto revisa trabajos anteriores en el área de explicabilidad en ciberseguridad, mencionando métodos como SHAP aplicados de manera no federada. Compara

indirectamente al mostrar cómo el método federado supera los desafíos de la centralización de datos y la falta de escalabilidad inherente a los métodos anteriores.

- Diferencias en la Escalabilidad y Distribución de Datos:

Mientras que los enfoques no federados a menudo enfrentan desafíos significativos en la centralización y procesamiento de grandes volúmenes de datos de seguridad distribuidos, el enfoque federado propuesto aprovecha la distribución de datos entre múltiples clientes. Esto no solo mejora la escalabilidad al reducir la carga en un servidor centralizado, sino que también preserva la privacidad al mantener los datos sensibles en los dispositivos locales.

- Interpretación y Consistencia de las Explicaciones:

A diferencia de los métodos no federados que pueden variar en la interpretación de explicaciones debido a diferencias en los conjuntos de datos y modelos locales, el enfoque federado propuesto utiliza técnicas como k-means federado para establecer una línea de base común de datos de fondo. Esto asegura que las explicaciones generadas por SHAP sean coherentes y comparables entre los diferentes clientes federados, facilitando así una interpretación más robusta y generalizable de las predicciones del modelo.

- Privacidad y Seguridad Mejoradas:

Los métodos no federados pueden comprometer la privacidad al requerir el intercambio de datos sensibles o la centralización de información crítica. En contraste, el enfoque federado minimiza estos riesgos al permitir que cada cliente mantenga el control sobre sus propios datos mientras contribuye de manera colaborativa al proceso de aprendizaje del modelo global. Esto no solo fortalece la seguridad de los datos, sino que también fomenta una mayor confianza por parte de los usuarios y las organizaciones en el uso de tecnologías avanzadas de ciberseguridad.

- Adaptabilidad y Eficiencia en el Contexto de FL:

A medida que el aprendizaje federado (federated learning FL, por sus siglas en inglés) se convierte en un estándar emergente en la ciberseguridad debido a sus ventajas en privacidad y eficiencia, el método propuesto demuestra cómo las técnicas de explicabilidad pueden adaptarse eficazmente a entornos distribuidos y heterogéneos. Esto asegura que las organizaciones puedan implementar soluciones de ciberseguridad robustas y efectivas sin comprometer la integridad de sus datos ni la efectividad de sus sistemas de detección de anomalías.

6.8 Impacto de las Mejoras

- **Propuestas y Taxonomías**: El método propuesto presenta una arquitectura novedosa que incluye el uso de k-means federado para generar un conjunto de datos de fondo y SHAP para explicar las decisiones del modelo. Esto podría considerarse una nueva propuesta en la aplicación de SHAP en entornos federados para ciberseguridad.
- **Resultados Relevantes:** Se mencionan métricas de validación como el índice CH para evaluar la calidad de los clusters de anomalías detectadas. Además, se discute cómo el método permite una interpretación más clara y distribuida de las anomalías en comparación con enfoques centralizados.

6.9 Resultados relevantes

Reducción de la Transmisión de Datos: Un beneficio práctico del método propuesto es que todas las etapas federadas se pueden realizar en una sola comunicación, reduciendo así la transmisión de datos entre los clientes y el servidor de agregación de FL. Esto mejora la eficiencia de comunicación en comparación con enfoques que requieren múltiples rondas de intercambio de datos.

Validación de Clústeres Anómalos: Se adapta la métrica de validación interna Calinski-Harabasz para entornos distribuidos, permitiendo estimar el número adecuado de clústeres anómalos entre todos los clientes. Esta validación es crucial para garantizar la robustez de la identificación y caracterización de las anomalías detectadas.

Tabla 20. Características de la Explicabilidad Federada para la Caracterización de Anomalías en Redes

Aspecto	Descripción
Mejoras en la Explicabilidad	Uso de Kernel SHAP en un entorno federado para mejorar la explicabilidad de modelos de detección de anomalías en redes. Método innovador en aplicar SHAP en contextos distribuidos y heterogéneos.
Marco k-means federado	Propuesta de un marco que utiliza k-means federado para establecer una línea de base común de datos de fondo y SHAP para explicar decisiones de modelo. Enfoque adaptado para entornos federados en ciberseguridad.

Comparación con Enfoques Anteriores	Contraste directo con métodos no federados que enfrentan desafíos de centralización de datos y escalabilidad al depender de infraestructuras centralizadas para el procesamiento y almacenamiento de grandes volúmenes de datos. En contraste, el enfoque federado aprovecha la distribución de datos entre múltiples clientes, permitiendo que cada nodo local mantenga el control sobre sus datos mientras contribuye de manera colaborativa al proceso de aprendizaje. Esto no solo mejora la escalabilidad al distribuir la carga de trabajo, sino que también fortalece la privacidad al reducir la necesidad de compartir datos sensibles de manera centralizada.
Impacto de las Mejoras	Reducción significativa de la transmisión de datos al realizar todas las etapas federadas en una sola comunicación, mejorando la eficiencia de comunicación comparado con enfoques que requieren múltiples rondas de intercambio de datos. Adaptación de métricas como el índice CH para validar clusters anómalos en entornos distribuidos, asegurando la robustez en la identificación de anomalías.
Propuestas y Taxonomías	Introducción de una arquitectura novedosa que incluye el uso de k-means federado para generar un conjunto de datos de fondo y SHAP para explicar decisiones del modelo en ciberseguridad.
Resultados Relevantes	Identificación exitosa de anomalías en datasets evaluados, asignación de etiquetas para caracterizar grupos de anomalías, facilitando alertas contextualizadas interoperables con herramientas de terceros.

7. Equidad Contrafactual Contrastiva en la Toma de Decisiones Algorítmicas (Mutlu, Yousefi, & Garibay, 2022).

Aunque existen numerosas definiciones de equidad a través del razonamiento causal, como la discriminación no resuelta, la no discriminación por proxy y la inferencia justa, la noción más comúnmente utilizada es la equidad contrafactual, que formula la equidad como la equivalencia entre los datos reales y su cantidad contrafactual. Recientemente, Zennaro et al.(2018) también propusieron otro concepto de equidad algorítmica aplicable a marcos de inferencia causal, llamado equidad contrastiva. En términos generales, las explicaciones contrafactuales buscan responder la pregunta "¿qué pasaría si?", mientras que los marcos contrastivos se enfocan en la pregunta "¿por qué esto y no aquello?".

7.1 Explicaciones Contrastivas y Contrafactuales en el Juicio Causal

Los contrafactuales son ampliamente utilizados en la ciencia cognitiva como mecanismos explicativos. En esencia, suponen que la mente compara eventos reales con escenarios alternativos, creando representaciones mentales de dichos escenarios . Los contrafactuales describen eventos y estados del mundo que no ocurrieron, contradiciendo hechos globales de manera implícita o explícita . Un contrafactual es una declaración condicional sobre una posibilidad alternativa y sus consecuencias, como "¿Cuál sería el resultado (si Y ocurre o no) si P ocurriera en lugar de Q?" .

Por otro lado, la investigación en humanidades y ciencias sociales indica que las explicaciones son inherentemente contrastivas . En el razonamiento contrastivo, una explicación busca responder la pregunta del porqué en relación con la causa de un evento, considerando alternativas hipotéticas no ocurridas, por ejemplo, "¿Por qué sucedió Y en lugar de Z?" . Se argumenta que las explicaciones contrastivas son capaces de diferenciar respuestas a preguntas explicativas mediante el uso de una amplia gama de alternativas contrastivas hipotéticas, proporcionando detalles suficientes sobre el razonamiento detrás de la pregunta .

Aunque las explicaciones contrafactuales y contrastivas parecen similares, difieren en ciertos aspectos. Lombrozo (2012) sostiene que ambos tipos de razonamiento consideran casos no ocurridos en comparación con la realidad, y que el razonamiento contrafactual es un subconjunto del razonamiento contrastivo . En el razonamiento contrafactual, se enfatiza la necesidad de un factor (por ejemplo, "¿Habría fallado el empleado si no fuera mujer?"), mientras que en el razonamiento contrastivo se enfoca en la suficiencia de un atributo, comparando efectos ausentes (por ejemplo, "¿Qué hizo la diferencia entre el empleado que falló y los que no fallaron?") .

7.2 Explicaciones Contrastivas y Contrafactuales en el Contexto de la Equidad

En años recientes, el razonamiento contrafactual ha sido adoptado en la literatura sobre equidad, inicialmente introducido por Kusner et al.(2017) .Matemáticamente, dado el atributo sensible A y la variable observada X, un predictor $\hat{Y}$ es "contrafactualmente" justo si cumple con:

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$$

para todos y y $a \neq a'$.

Esta fórmula garantiza que la distribución de las predicciones sea la misma en el mundo real y en un mundo contrafactual donde los atributos sensibles cambian, manteniendo constantes todas las demás condiciones.

Aunque la equidad contrafactual aborda la equidad para personas de diferentes grupos, este criterio sólo se aplica a nivel de población. Las categorías sociales a menudo no permiten la manipulación de contrafactuales, lo que dificulta la evaluación de la equidad . Para superar este reto, Chakrabarti et al.(2020) propusieron la equidad contrastiva en la toma de decisiones algorítmicas , sugiriendo tres definiciones de equidad contrastiva.

La primera definición busca responder "¿Es justo tomar la decisión D para el individuo I, en lugar de la decisión D′?" Dado el atributo sensible A y las variables observadas X para un individuo iii, un predictor Y^ es justo en términos contrastivos D si:

$$P(\hat{Y}_{A_i\leftarrow a}(U_i) = y \mid X_i = x, A_i = a) = P(\hat{Y}_{A_i\leftarrow a'}(U_i) = y \mid X_i = x, A_i = a)$$

La segunda definición aborda la equidad en términos de individuos diferentes, respondiendo "¿Es justo tomar la decisión D para el individuo I, pero D′ para el individuo J?" Aquí, un predictor Y^ es justo en términos contrastivos I si:

$$P(\hat{Y}_{A_i\leftarrow a_i}(U_i) = y \mid X_i = x, A_i = a_i) = P(\hat{Y}_{A_i\leftarrow a_i'}(U_i) = y \mid X_i = x, A_i = a_i)$$

La tercera definición examina la equidad en el tiempo, preguntando "¿Es justo tomar la decisión D para el individuo I en el tiempo t, pero D′ en el tiempo t′?" La cuestión principal es si es justo que una decisión para un individuo cambie con el tiempo. Así, un predictor Y^ es justo en términos contrastivos T si:

$$P(\hat{Y}_{A_i \leftarrow a_i}(U_i) = y \mid X_i = x_i(t), A_i = a_i) = P(\hat{Y}_{A_i \leftarrow a_i'}(U_i) = y \mid X_i = x_i(t), A_i = a_i)$$

7.3 Propuesta del trabajo

El estudio propone una nueva definición de equidad algorítmica denominada "equidad contrastiva contrafactual", que integra dos técnicas de razonamiento ampliamente utilizadas en el análisis explicativo de la toma de decisiones de IA: el razonamiento contrafactual y el contrastivo. Estos enfoques se consideran complementarios entre sí y pueden explicar detalladamente el razonamiento detrás de una pregunta al considerar un conjunto amplio de alternativas hipotéticas contrastivas (contrafactuales). Dado que el razonamiento contrastivo contrafactual se adopta ampliamente en la literatura de IA explicativa, se cree que esta definición satisface un tratamiento equitativo completo entre dos individuos.

Para probar la eficiencia de la definición de equidad propuesta, se realizaron experimentos en dos conjuntos de datos relacionados con la equidad más utilizados: el conjunto de datos UCI Adult[2] y el conjunto de datos de crédito alemán[3], aplicando una técnica de aumento de datos para mejorar la equidad del algoritmo de toma de decisiones tanto para grupos protegidos como no protegidos. Como algoritmos de predicción, se utilizaron regresión logística (LR) con algoritmo de optimización BFGS de memoria limitada, perceptrón multicapa (MLP) y máquina de vectores de soporte (SVM).

A pesar de la existencia de múltiples atributos sensibles en ambos conjuntos de datos, se seleccionó un atributo sensible en el análisis experimental por simplicidad. Como criterios de desempeño, no solo se midieron los cambios en la precisión del análisis de predicción, sino que también se calcularon medidas de equidad como la exactitud, las probabilidades igualadas y la paridad demográfica. Por lo tanto, se puede argumentar que los resultados finales muestran la eficiencia de los resultados propuestos en términos del equilibrio entre precisión y equidad.

Figura 9. Exactitud de los Algoritmos de Predicción utilizando el Conjunto de Datos UCI Adult.

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Método	LR	MLP	SVM
Sin mitigación	0.7633	0.7775	0.7598
Contrafactual	0.7511	0.7545	0.7733
Contrafactual Contrastivo	0.7539	0.7604	0.7721

En el conjunto de datos Adulto, el objetivo fue clasificar si el ingreso anual de un individuo supera los 50,000 dólares basándose en datos del censo, desbiasando el resultado con respecto a los atributos de sexo. A pesar de la principal preocupación por mejorar la equidad del algoritmo de clasificación, también se apuntó a no comprometer la precisión. Por lo tanto, se calcularon inicialmente los valores de exactitud de tres algoritmos de clasificación diferentes: primero, sin modificación en el conjunto de datos y el método de clasificación (sin mitigación); segundo, cuando se aumentaron los datos contrafactuales del atributo sexo antes de la clasificación (contrafactual); y tercero, cuando se aumentaron los datos contrafactuales de ejemplos contrastivos (contrafactual contrastivo). Aunque los criterios de equidad adicionales en la optimización de una estrategia de clasificación afectan su precisión, la técnica de aumento de datos contrafactual parece no dañar significativamente el rendimiento general. En la implementación del clasificador SVM, incluso se observaron valores de precisión más altos. Por otro lado, el aumento de datos contrastivo contrafactual produjo resultados ligeramente mejores que el aumento de datos contrafactuales en la implementación de LR y MLP. Estos resultados muestran que la estrategia propuesta no compromete significativamente la precisión del algoritmo de clasificación mientras intenta mitigar el sesgo en su implementación. Además, el conjunto de datos de crédito alemán también proporcionó resultados similares.

Para comparar y contrastar el desempeño de equidad de la definición de equidad propuesta, se calcularon los puntajes de exactitud, probabilidades igualadas y paridad demográfica de los algoritmos de clasificación LR, MLP y SVM cuando no se realizó mitigación, solo se realizó el aumento de datos contrafactuales y se aplicó la técnica de aumento de datos contrafactual contrastivo (ver Figura 9).

En el caso del conjunto de datos Adulto de UCI, dado que los algoritmos tienden a asignar erróneamente valores de ingresos bajos a las mujeres en comparación con los hombres, pero predicen valores verdaderos de manera similar entre diferentes atributos de sexo, los resultados verdaderos (tanto TP como TN) no se ven muy afectados. Por lo tanto, los resultados de paridad de exactitud muestran diferentes patrones en la utilización de técnicas de mitigación de sesgo. Mientras que la paridad de exactitud disminuye, es decir, se obtienen valores de precisión similares para individuos en ambos grupos protegidos y no protegidos, en la utilización de datos de aumento contrafactual antes de la implementación de LR, el

valor aumentó en la aplicación de MLP y SVM. El aumento de datos contrastivo contrafactual, por otro lado, produjo mejores resultados con SVM, pero resultados menos óptimos con LR y MLP. Además, en el conjunto de datos de crédito alemán, se obtuvieron las mejores actuaciones con el aumento de datos contrastivo contrafactual antes de todos los algoritmos de clasificación, seguido por el aumento de datos contrafactuales.

En cuanto a la igualdad de oportunidades, dado que calcula las diferencias de valores TP + TN para individuos femeninos y masculinos, el aumento de datos contrafactual no produce mejores resultados en todos los casos. Por ejemplo, proporciona puntajes de paridad más favorables en la aplicación de MLP y SVM en el conjunto de datos Adulto de UCI y en el de LR y SVM en el conjunto de datos de crédito alemán en comparación con el caso en que no se realiza mitigación de sesgo. Mientras tanto, el aumento de datos contrastivo contrafactual produce los mejores desempeños con los puntajes de igualdad de oportunidades más bajos en la implementación de las tres técnicas de clasificación en ambos conjuntos de datos. En la comparación de paridad demográfica, también se observan resultados similares. El aumento de datos contrastivo contrafactual produce los mejores resultados en ambos conjuntos de datos independientemente de la técnica de clasificación implementada.

Esto se puede explicar de la siguiente manera: El aumento de datos contrafactuales simplemente duplica los datos y convierte el atributo de sexo para evitar resultados sesgados hacia este atributo sensible. Sin embargo, las proporciones de predicciones verdaderas y falsas no son las mismas para mujeres y hombres, y agregar todos los datos contrafactuales cambiará la dirección del sesgo, pero no lo eliminará completamente. Por otro lado, el aumento de datos contrastivo contrafactual proporciona un preprocesamiento de datos más lógico, ya que los contrafactuales (datos en los que el atributo de sexo se convierte solo) se aumentan solo para ejemplos contrastivos. En esta técnica, el puntaje de necesidad en la Ecuación:

$$NEC_a^{a'}(U) = \mathbb{P}(\hat{Y}_{A\leftarrow a'}(U) = y' | \mathcal{X} = x, A = a, Y = y)$$

mide el porcentaje de decisiones positivas del algoritmo que son atribuibles o debidas al valor del atributo *x*, mientras que el puntaje de suficiencia en la Ecuación:

$$SUF_a^{a'}(U) = \mathbb{P}(\hat{Y}_{A\leftarrow a}(U) = y | \mathcal{X} = x, A = a', Y = y')$$

relaciona que "¿Cuál sería la probabilidad de que para individuos con atributos *x*, la decisión del algoritmo fuera positiva en lugar de negativa si *A* hubiera sido *a* en lugar de *a*'? En el aumento de datos contrastivo contrafactual, concatenamos ejemplos contrafactuales de i)

mujeres si el resultado es realmente negativo y el resultado del algoritmo de predicción con resultado negativo; sin embargo, el algoritmo predeciría positivo si el individuo fuera hombre, y ii) hombres si el resultado es positivo y el resultado del algoritmo de predicción con resultado positivo; sin embargo, el algoritmo predeciría negativo si el individuo fuera mujer. Este aumento selectivo de datos contrastivo y contrafactual produce un conjunto de datos más equitativo en el preprocesamiento de mitigación de sesgo.

Los datos experimentales muestran que la técnica de aumento reduce significativamente el sesgo en términos de tres criterios de equidad: paridad de exactitud, paridad de probabilidades igualadas y paridad demográfica entre grupos protegidos y no protegidos, mientras mantiene casi constante la precisión.

7.4 Avances Recientes en Explicabilidad de Modelos 'Caja Negra'.

En los estudios recientes, se ha introducido la noción de "equidad contrastiva contrafactual" como una nueva definición en la literatura de AI explicativa. Esta metodología combina el razonamiento contrafactual y el contrastivo para mejorar la equidad en algoritmos de toma de decisiones automáticas (ver Tabla 22). A diferencia de los métodos tradicionales que se centran principalmente en la equidad contrafactual, la equidad contrastiva permite evaluar decisiones algorítmicas no sólo en términos de diferentes grupos, sino también en términos de diferentes decisiones para un mismo individuo en diferentes circunstancias.

Esta innovación ha propuesto una nueva taxonomía para evaluar la equidad en algoritmos de decisión automatizados, ofreciendo tres definiciones distintas de equidad contrastiva que abordan diferentes aspectos de la toma de decisiones algorítmicas. Experimentos con conjuntos de datos reales como UCI Adult y crédito alemán han demostrado que técnicas como el aumento de datos contrafactual contrastivo pueden reducir significativamente el sesgo en términos de paridad de precisión, probabilidades igualadas y paridad demográfica, sin comprometer en gran medida la precisión general de los modelos de clasificación (LR, MLP, SVM).

Comparando con enfoques anteriores, se observa que el aumento de datos contrafactual contrastivo produce los mejores resultados en términos de equidad en ambos conjuntos de datos evaluados, superando a otros métodos de mitigación de sesgo como el aumento de datos contrafactual estándar. La validación empírica de estas mejoras ha mostrado mejoras significativas en la equidad medida por diversos criterios, respaldando la robustez teórica y la efectividad práctica de la equidad contrastiva contrafactual.

Impacto de las Mejoras:

- **Propuestas y Taxonomías**: La introducción de la equidad contrastiva contrafactual ha propuesto una nueva taxonomía para evaluar la equidad en algoritmos de decisión automatizados. Esta taxonomía ofrece tres definiciones distintas de equidad contrastiva, abordando diferentes aspectos de la toma de decisiones algorítmicas.
- **Resultados Relevantes**: Los estudios de caso utilizando conjuntos de datos reales (como UCI Adult y crédito alemán) han demostrado que la implementación de técnicas como el aumento de datos contrafactual contrastivo puede reducir significativamente el sesgo en términos de paridad de precisión, probabilidades igualadas y paridad demográfica, sin comprometer en gran medida la precisión general de los modelos de clasificación (LR, MLP, SVM).

Comparación y Evaluación de Resultados:

- **Comparación de Desempeño**: Se observó que el aumento de datos contrafactual contrastivo produjo los mejores resultados en términos de equidad en ambos conjuntos de datos evaluados, superando a otros métodos de mitigación de sesgo como el aumento de datos contrafactual estándar.
- **Validación Empírica**: Los experimentos mostraron mejoras significativas en la equidad medida por diversos criterios, lo que sugiere que la equidad contrastiva contrafactual no sólo es teóricamente sólida, sino también efectiva en la práctica.

Tabla 22. Avances en Explicabilidad y Equidad en Modelos de AI: Equidad Contrastiva Contrafactual.

Aspecto	Descripción
Mejoras en la Explicabilidad	Se introduce la noción de "equidad contrastiva contrafactual" como una nueva definición en la literatura de AI explicativa. Esta metodología combina el razonamiento contrafactual y el contrastivo para mejorar la equidad en algoritmos de toma de decisiones automáticas.
Marco	• Avances Recientes: Introducción de la equidad contrastiva contrafactual en la explicación de modelos de 'caja negra'. • Método Innovador: Integración de razonamiento contrafactual y contrastivo para abordar sesgos algorítmicos y mejorar la comprensión de decisiones automatizadas.

Comparación con Enfoques Anteriores	• Comparación con Métodos Tradicionales: Enfoques anteriores se centraban principalmente en la equidad contrafactual, mientras que el enfoque actual amplía la evaluación a través de la equidad contrastiva, considerando múltiples perspectivas de justicia algorítmica.
Impacto de las Mejoras	• Propuestas y Taxonomías: Introducción de la equidad contrastiva contrafactual como nueva definición en el campo de la equidad algorítmica, proporcionando un marco más completo para evaluar la equidad en sistemas de inteligencia artificial. • Resultados Relevantes: Experimentos muestran una reducción significativa del sesgo algorítmico al aplicar técnicas de aumento de datos contrafactual contrastivo, manteniendo altos niveles de precisión en la clasificación.
Propuestas y Taxonomías	• Definición Propuesta: Equidad contrastiva contrafactual, que aborda los desafíos de la equidad al considerar diferentes contextos y grupos afectados por decisiones algorítmicas. • Taxonomías Presentadas: Tres definiciones de equidad contrastiva que exploran la equidad individual, grupal y temporal en la toma de decisiones automatizadas.
Resultados Relevantes	• Casos de Estudio: Uso de conjuntos de datos reales como UCI Adult y crédito alemán para demostrar la eficacia de las mejoras propuestas. • Métricas Evaluadas: Paridad de precisión, probabilidades igualadas y paridad demográfica muestran mejoras significativas al mitigar el sesgo algorítmico mediante estrategias avanzadas de explicabilidad.

8. Sistemas Explicadores Basados en Reglas Difusas para Redes Neuronales Profundas: De la Explicabilidad Local a la Comprensión Global (Aghaeipoor, Sabokrou, & Fernández, 2023).

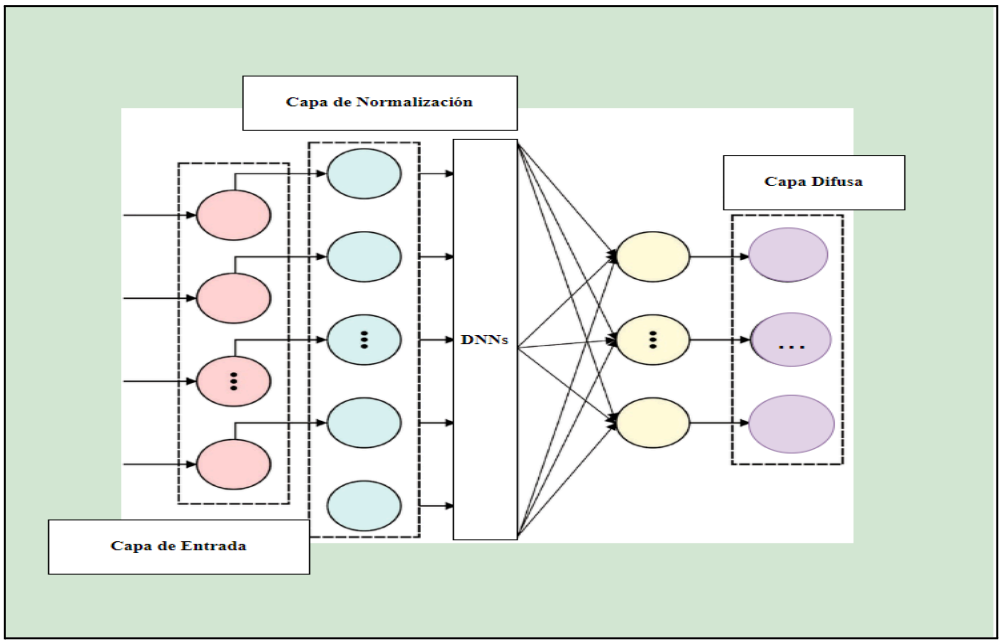
Este artículo presenta los Sistemas de Explicación Basados en Reglas Difusas (FRBES), una técnica para explicar redes neuronales profundas (DNNs). Los FRBES utilizan reglas difusas

para imitar el comportamiento de las DNNs, proporcionando explicaciones más comprensibles que los métodos tradicionales.

En particular, la Tabla 23 resume los avances y aplicaciones de los FRBES en la interpretación y explicación del comportamiento de modelos de DNNs, destacando mejoras significativas en la interpretabilidad y la precisión (véase Tabla 23).

La **Fig.10** muestra el flujo de datos a través de varias etapas dentro de un Sistema de Explicación Basado en Reglas Difusas (FRBES). Este flujo incluye la normalización inicial de datos, el procesamiento mediante una DNN, y finalmente, la aplicación de lógica difusa para generar explicaciones interpretables del modelo.

Figura 10. Estructura del Sistema de Explicación basado en Reglas Difusas (FRBES). (Elaboración propia).



8.1 Proceso de Creación de un FRBES.

1. Entrenamiento de la DNN y Obtención de la Importancia de las Características:
    - **Entrenamiento de la DNN:**
        - Se entrena una DNN con tres capas ocultas y función de activación ReLU en un conjunto de datos específico.
        - Se exploran diferentes hiperparámetros (número de neuronas por capa, tamaño del lote y tasa de aprendizaje) utilizando búsqueda en cuadrícula.

        - Se utiliza el optimizador Adam para minimizar la pérdida de entropía cruzada durante 150 épocas.
    - **Importancia de las Características:**
        - Se determina la importancia de las características utilizando algoritmos de atribución como DeepLIFT, Gradientes Integrados y Gradient SHAP. Estos métodos permiten entender qué características tienen mayor impacto en las predicciones de la DNN.
2. Extracción y Optimización de Reglas Difusas:
    - **Generación de la Base de Datos (DB):**
        - Se utilizan funciones de membresía triangulares y particionamiento uniforme difuso para transformar los valores concretos en grados difusos.
    - **Generación de la Base de Reglas (RB):**
        - Se crean reglas iniciales utilizando un enfoque adaptado del algoritmo de Chi. Estas reglas son optimizadas para mejorar la interpretabilidad y el rendimiento predictivo del modelo.
    - **Optimización de las Reglas:**
        - **Poda de Reglas Redundantes:** Las reglas redundantes se eliminan para asegurar que las reglas finales sean las más informativas y de menor longitud.
        - **Selección de las Mejores Reglas:** Se retienen las reglas más confiables, clasificadas según sus valores de confianza, manteniendo solo el α% superior de cada clase.

8.2 Ventajas de los FRBES

- **Interpretabilidad:** Las reglas difusas son más fáciles de entender que los modelos de DNN tradicionales, facilitando la comprensión del proceso de toma de decisiones.
- **Fidelidad:** Los FRBES pueden imitar el rendimiento de las DNN originales con alta precisión.
- **Compacidad:** Los FRBES utilizan un número reducido de reglas y características, haciéndolos más manejables y eficientes.

8.3 Comparación con Enfoques Anteriores

- **ECLAIRE:** El artículo compara FRBES con ECLAIRE, un algoritmo de última generación para la extracción de reglas de las DNNs. FRBES supera a ECLAIRE en términos de interpretabilidad, fidelidad y compacidad.

8.4 Resultados Experimentales

- **Conjuntos de Datos:** Se utilizaron seis conjuntos de datos de clasificación de diferentes áreas, aplicando validación cruzada de cinco pliegues.
- **Criterios de Evaluación:** Se evaluaron precisión (ACC), área bajo la curva (AUC), fidelidad (Fidelity), número de características (#Fc), número de reglas (#R), longitud promedio de las reglas (ARL) y tiempo de procesamiento.
- **Pruebas Estadísticas:** Se realizaron pruebas estadísticas, como las de Friedman y Holm, para comparar el rendimiento de los métodos con un nivel de significancia de $\alpha = 0.1$.
- **Métodos Comparativos:** Se compararon los FRBES con ECLAIRE y Chi_FRBCS, mostrando que los FRBES logran una alta fidelidad y precisión, manteniendo una complejidad significativamente menor.

8.5 Clasificación de FRBES

- **Escenario:**
  - **Post-hoc:** FRBES se aplica a un modelo DNN ya entrenado.
  - **Modelo específico:** FRBES está diseñado específicamente para explicar DNNs.
- **Alcance:**
  - **Local y Global:** FRBES proporciona explicaciones tanto locales como globales.
- **Tipo de Problema:**
  - **Clasificación:** El artículo se centra en la clasificación.
- **Dato de Entrada:**
  - **Numérico/Categórico:** FRBES se aplica a datos tabulares.
- **Formato de Salida:**
  - **Reglas Difusas:** FRBES genera un conjunto de reglas difusas que explican el comportamiento del modelo.

8.6 Impacto Práctico

Los resultados muestran que FRBES puede ser una herramienta útil para mejorar la explicabilidad de las DNNs en aplicaciones del mundo real, especialmente en áreas como la medicina donde la interpretabilidad es crucial.

Experimentos y Resultados

1. **Configuración Experimental:**
   - Se utilizaron seis conjuntos de datos de clasificación de diferentes áreas.

- La DNN se configuró con tres capas ocultas y función de activación ReLU, explorando diferentes hiperparámetros mediante búsqueda en cuadrícula.
- Se entrenaron redes utilizando el optimizador Adam durante 150 épocas.
- FRBES se construyó sobre tres métodos de atribución: DeepLIFT, Gradientes Integrados y Gradient SHAP, además de dos versiones agregadas (AGG-Mean y AGG-Var).

2. **Evaluación y Pruebas Estadísticas:**
   - Se reportaron valores de precisión, AUC, fidelidad, número de características, número de reglas, longitud promedio de las reglas y tiempo de procesamiento.
   - La fidelidad mide qué tan bien las explicaciones reflejan el comportamiento de la DNN original.
   - Se realizaron pruebas estadísticas para comparar el rendimiento de los métodos.

Los FRBES se construyen entrenando una DNN, extrayendo reglas difusas a partir de las características importantes y optimizándolas para obtener las reglas más informativas y confiables, proporcionando así explicaciones interpretables y precisas para el comportamiento de las DNNs.

Tabla 23. Avances y Aplicaciones de Sistemas de Explicación Basados en Reglas Difusas (FRBES) para Redes Neuronales Profundas.

Aspecto	Descripción
Mejoras en la Explicabilidad	Introduce mejoras significativas en la capacidad de entender y explicar el funcionamiento interno de modelos complejos como las DNNs, utilizando reglas difusas y métodos de atribución.
Marco	Se basa en el uso de Sistemas de Explicación Basados en Reglas Difusas (FRBES) para proporcionar explicaciones interpretables y precisas del comportamiento de las DNNs
Comparación con Enfoques Anteriores	Se contrasta con métodos anteriores como ECLAIRE, demostrando una mejora en interpretabilidad, fidelidad y eficiencia en la explicación de modelos DNN.
Impacto de las Mejoras	Destaca cómo las mejoras en la

	explicabilidad benefician aplicaciones prácticas, como en medicina, al facilitar la toma de decisiones basadas en modelos complejos.
Propuestas y Taxonomías	Propone nuevas taxonomías o categorizaciones para entender y clasificar métodos de explicabilidad como los FRBES, enriqueciendo el campo de la IA explicativa.
Resultados Relevantes	Presenta resultados experimentales que validan la efectividad de los FRBES en términos de precisión, interpretabilidad y eficiencia comparativa con métodos tradicionales y recientes.

9. Inteligencia Artificial Explicable (XAI): Lo que sabemos y lo que queda por lograr para una Inteligencia Artificial Confiable (Ali et al., 2023).

El artículo explora la explicabilidad en la inteligencia artificial desde tres perspectivas distintas: usuario, aplicación y gobierno. Se enfoca en analizar cuatro ejes que son fundamentales para proporcionar explicaciones efectivas en sistemas inteligentes. Estos ejes abarcan *la explicabilidad de datos, la explicabilidad del modelo, la explicabilidad post-hoc y la evaluación de las explicaciones*. Además, se examinan principios clave como equidad, seguridad, responsabilidad, ética y privacidad para fortalecer la confianza del usuario en XAI.

La Tabla 27 aborda la comprensión y confianza en los sistemas inteligentes a través de un marco que considera las perspectivas del usuario, el de la aplicación y la del gobierno, teniendo en cuenta también la evolución de los enfoques, proponiendo una metodología organizada en cuatro ejes para mejorar la interpretabilidad y precisión de las explicaciones proporcionadas.

Tabla 27.Aspectos Clave del paper 'Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence.

Aspecto	Descripción
Mejoras en la Explicabilidad	Proporciona un marco estructurado que permite comprender cómo la explicabilidad se aplica a diferentes partes del ecosistema

	de inteligencia artificial.
Marco	El artículo explora la explicabilidad en la inteligencia artificial desde tres perspectivas: usuario, aplicación y gobierno. Se centra en analizar cuatro ejes fundamentales para proporcionar explicaciones efectivas en sistemas inteligentes.
Comparación con Enfoques Anteriores	Se distinguen modelos de caja blanca, caja gris y caja negra, evaluando cómo la XAI puede lograr modelos confiables con un equilibrio óptimo entre interpretabilidad y precisión.
Impacto de las Mejoras	Se subraya la importancia de principios como equidad, seguridad, responsabilidad, ética y privacidad para fortalecer la confianza del usuario en XAI.
Propuestas y Taxonomías	Introduce una metodología organizada en cuatro ejes: explicabilidad de datos, del modelo, post-hoc y evaluación de explicaciones.
Resultados Relevantes	Destaca la relación entre explicabilidad e interpretabilidad como requisito previo para explicaciones efectivas, enfatizando la interactividad para verificar la fiabilidad del modelo.

10. Explicaciones Lógicas Basadas en Entropía de Redes Neuronales. (Barbiero et al., 2022).

El artículo "Entropy-Based Logic Explanations of Neural Networks" (Barbiero et al., 2022) propone un enfoque para proporcionar explicaciones lógicas en redes neuronales mediante una capa lineal basada en entropía.

Esta nueva capa introducida en las redes neuronales permite explicaciones lógicas basadas en entropía al tomar como entrada conceptos del espacio C.

Durante el cálculo de esta capa, se obtienen dos resultados principales:

- Embebidos h^i: Resultados de la capa lineal, similares a los obtenidos en cualquier capa lineal estándar.
- Tabla de verdad T^i: Explica cómo la red utilizó los conceptos para hacer predicciones para la clase objetivo i, representando cómo estos conceptos se relacionan con las predicciones de la red.

Cada clase del problema requiere una capa basada en entropía independiente, lo que implica que se instancia una capa separada para cada clase que se desea predecir.

El enfoque se centra en inferir para observaciones individuales, representadas por una tupla de conceptos $c \in C$. La red neuronal f^i predice la membresía de clase para la clase i-ésima del problema.

En problemas con múltiples clases, se emplean múltiples instancias de esta capa, una para cada clase objetivo, lo que facilita que cada clase disponga de su propio método de explicación basado en entropía.

Aunque cada clase requiere una capa basada en entropía independiente, compartir las capas ocultas entre las redes de distintas clases ayuda a optimizar el uso de recursos computacionales.

Para garantizar que las explicaciones sean simples y comprensibles, se utilizan fórmulas lógicas de primer orden (FOL) derivadas de las tablas de verdad. Estas fórmulas muestran de manera clara cómo se relacionan los conceptos de entrada con las predicciones realizadas por la red neuronal.

En la Tabla 28 aborda la explicabilidad a través de un marco que considera las perspectivas del usuario, el de la aplicación y la del gobierno, teniendo en cuenta también la evolución de los enfoques, proponiendo una metodología organizada en cuatro ejes para mejorar la interpretabilidad y precisión de las explicaciones proporcionadas.

Tabla 28. Aspectos Clave en la Explicabilidad de la Inteligencia Artificial según el Artículo

Aspecto	Descripción
Mejoras en la Explicabilidad	Introduce una capa lineal basada en entropía para explicaciones lógicas en redes neuronales.
Marco	Basado en conceptos del espacio C (espacio C puede ser cualquier conjunto de características, atributos o variables que son

	relevantes para el problema que se está abordando).
Comparación con Enfoques Anteriores	Decision Trees: modelo de aprendizaje automático intrínsecamente interpretable que proporciona explicaciones en forma de reglas lógicas. Bayesian Rule Lists (BRL): Método de minería de reglas que utiliza estadísticas bayesianas para extraer un conjunto óptimo de reglas, proporcionando explicaciones en forma de reglas lógicas. ψ Networks: Redes neuronales explicables que utilizan capas simbólicas intermedias para generar fórmulas de lógica de primer orden, permitiendo entender el proceso de decisión del modelo.
Impacto de las Mejoras	Mejora la comprensión del razonamiento lógico de las redes neuronales.
Propuestas y Taxonomías	Proporciona fórmulas lógicas de primer orden derivadas de tablas de verdad.
Resultados Relevantes	Embebidos h^i y Tabla de verdad T^i que explican cómo se relacionan los conceptos con las predicciones de la red.

11. Redes explicadas por Lógica (Ciravegna et al., 2023).

Una solución posible para proporcionar explicaciones entendibles por humanos es utilizar un lenguaje formal expresivo y relacionado con el razonamiento, como la Lógica de Primer Orden (FOL). Las explicaciones FOL son declaraciones rigurosas y unívocas que pueden describir relaciones complejas entre conceptos. A diferencia de otras técnicas basadas en conceptos, las explicaciones basadas en lógica permiten una medición cuantitativa de su corrección y completitud, además de ser versátiles y simplificables. LENs pueden ser configurados de diversas maneras según las necesidades del usuario y las características del problema, siendo útiles tanto para clasificación supervisada como para configuraciones no supervisadas.

A continuación se presenta la Tabla 29, que detalla los aspectos clave discutidos en el artículo y sus implicaciones para la práctica y la investigación en inteligencia artificial.

Tabla 29.Aspectos Clave en la Explicabilidad de la Inteligencia Artificial según el Artículo.

Aspecto	Descripción
Mejoras en la Explicabilidad	Introduce un enfoque que combina redes neuronales con lógica formal, mejorando significativamente la explicabilidad de modelos de caja negra.
Marco	Propone una familia de modelos interpretables de aprendizaje profundo llamados Logic Explained Networks (LENs), que utilizan fórmulas simples de Lógica de Primer Orden (FOL) para proporcionar explicaciones comprensibles.
Comparación con Enfoques Anteriores	Se compara con modelos de caja blanca establecidos como árboles de decisión y listas de reglas bayesianas, demostrando que los LENs pueden producir clasificaciones superiores y explicaciones más completas.
Impacto de las Mejoras	Mejora la confianza en los modelos de inteligencia artificial al proporcionar explicaciones claras y razonamientos transparentes. Facilita una comprensión más profunda de cómo las decisiones del modelo se derivan de los datos, crucial para aplicaciones críticas y decisiones automatizadas.
Propuestas y Taxonomías	Propone un marco para integrar la lógica explicativa en redes neuronales, explorando diferentes configuraciones y aplicaciones en aprendizaje supervisado y no supervisado.
Resultados Relevantes	Resultados experimentales muestran que los LENs superan a modelos de caja blanca en términos de precisión y explicabilidad en diversas tareas y conjuntos de datos. Además, destacan su capacidad para mejorar la interpretación de modelos complejos.

Funcionamiento de las Redes Explicadas por Lógica (LENs).

Entradas y Datos de Entrada:

- Los LENs reciben datos de entrada, que pueden ser características interpretables como conceptos específicos (por ejemplo, atributos de pacientes en un conjunto de datos médico).

Proceso de Clasificación:

- Utilizan una arquitectura de red neuronal, como una CNN (Convolutional Neural Network) u otra estructura adecuada para el problema, para realizar la clasificación de los datos de entrada en categorías predefinidas.

Generación de Explicaciones:

- Además de clasificar, los LENs son diseñados para generar explicaciones en forma de lógica de primer orden (FOL). Estas explicaciones describen cómo se llegó a una decisión de clasificación específica a partir de los datos de entrada.

Formulación de Reglas Lógicas:

- Las reglas lógicas generadas por los LENs son claras, estructuradas y pueden expresar relaciones complejas entre los conceptos interpretativos (por ejemplo, "si una persona tiene fiebre y dolor de garganta, entonces es probable que tenga gripe").

Aplicaciones:

- Los LENs pueden ser utilizados tanto para clasificar datos de manera interpretable como para proporcionar explicaciones de modelos caja negra. Esto los hace útiles en aplicaciones donde se requiere transparencia y comprensión del proceso de toma de decisiones del modelo.

Flexibilidad y Adaptabilidad:

- La arquitectura de los LENs permite adaptarse a diferentes contextos y problemas, desde la clasificación directa hasta la explicación de modelos complejos, manteniendo la capacidad de generar explicaciones comprensibles para los usuarios finales.

12. Aprendizaje Profundo con Restricciones Lógicas (Giunchiglia, Stoian, y Lukasiewicz (2022)).

El paper "Deep Learning with Logical Constraints" analiza cómo integrar conocimientos previos expresados en lógica de primer orden (FOL) en modelos de aprendizaje profundo para mejorar su rendimiento y explicabilidad.

Los modelos neuronales han sido exitosos, pero a menudo son agnósticos al dominio y no aprovechan el conocimiento específico del problema.

Motivación y Objetivos

Los modelos neuronales han sido exitosos, pero a menudo no aprovechan el conocimiento específico del problema. Expresar este conocimiento en formas como ecuaciones algebraicas, restricciones lógicas o lenguaje natural puede:

- Mejorar el rendimiento.
- Permitir el aprendizaje con menos datos.
- Asegurar que las salidas del modelo sean consistentes con las restricciones lógicas.

Métodos para Integrar Restricciones Lógicas en Modelos de Aprendizaje Profundo.

El artículo presenta y compara diversos métodos para integrar restricciones lógicas, destacando su impacto en la interpretabilidad y rendimiento de los modelos.

En el contexto de la integración de restricciones lógicas en modelos de aprendizaje profundo, se han propuesto diversos métodos que han demostrado mejorar significativamente la interpretabilidad y rendimiento de los modelos (véase Tabla 30).

Métodos Basados en Pérdida

1. Regularización Basada en Semántica (SBR)
    - Incorpora restricciones lógicas mediante una regularización diferenciable, optimizando simultáneamente la precisión del modelo y la consistencia con reglas lógicas (Diligenti et al., 2012; Diligenti et al., 2017).
2. Destilación Iterativa del Conocimiento de Reglas
    - Transfiere conocimiento de reglas lógicas a través de un proceso de destilación, donde una red maestra guía el entrenamiento de una red estudiante (Hu et al., 2016a; Hu et al., 2016b).
3. Aprendizaje Abductivo (ABL)
    - Combina redes neuronales con componentes lógicos para ajustar las predicciones basadas en hechos generados que cumplen con el conocimiento lógico (Dai et al., 2019).

Métodos Basados en Estructuras Especializadas

1. Lógica Real y Redes Tensoriales Lógicas (LTNs)

    - Formaliza la lógica de primer orden en un contexto de números reales, integrando conocimiento lógico en la representación de datos (Serafini y d'Avila Garcez, 2016; Badreddine et al., 2022; Donadello et al., 2017; Marra et al., 2019).
2. Método Diferenciable Fuzzy Logics (DFL)
    - Traduce expresiones lógicas a funciones de pérdida diferenciables utilizando operadores difusos, facilitando la integración de restricciones lógicas en modelos de aprendizaje profundo (Van Krieken et al., 2020; Van Krieken et al., 2022).

Impacto en la Explicabilidad de Modelos 'Caja Negra'

Los avances en estos métodos han tenido un impacto significativo en la explicabilidad de los modelos de machine learning denominados 'caja negra':

- Propuestas: Han introducido nuevas técnicas para integrar restricciones lógicas en modelos complejos, mejorando la interpretación de resultados y la confianza en las decisiones del modelo.
- Taxonomías: Estos avances han contribuido a la creación de taxonomías más refinadas para clasificar enfoques de explicabilidad según cómo manejan las restricciones lógicas y su integración con la optimización de modelos.
- Resultados Relevantes: Se han logrado mejoras significativas en la consistencia y la interpretabilidad de los modelos, facilitando su aplicación en contextos donde la transparencia y la interpretación de decisiones son críticas.

Tabla 30. Integración de Restricciones Lógicas en Modelos de Deep Learning

Aspecto	Descripción
Mejoras en la Explicabilidad	Integración de restricciones lógicas para mejorar la interpretabilidad y rendimiento de modelos de deep learning.
Marco	Incorporación de conocimiento de dominio mediante restricciones lógicas en redes neuronales profundas.
Comparación con Enfoques Anteriores	Avances en la integración de restricciones lógicas comparados con enfoques tradicionales de redes neuronales.Los métodos revisados incluyen:

	Destilación Iterativa del Conocimiento de Reglas ○ Transfiere conocimiento de reglas lógicas a través de un proceso de destilación. **Aprendizaje Abductivo (ABL)** ○ Combina redes neuronales con componentes lógicos para ajustar las predicciones basadas en hechos generados que cumplen con el conocimiento lógico. **Lógica Real y Redes Tensoriales Lógicas (LTNs)** ○ Utiliza una formalización de lógica de primer orden en un contexto de números reales para integrar conocimiento lógico en la representación de datos. **Método Diferenciable Fuzzy Logics (DFL)** ○ Traduce expresiones lógicas a funciones de pérdida diferenciables utilizando operadores difusos.
Impacto de las Mejoras	Mejora significativa en la interpretación de modelos y garantía de cumplimiento de restricciones de conocimiento.
Propuestas y Taxonomías	Propuesta de una taxonomía para clasificar métodos basados en el tipo y complejidad de las restricciones lógicas.

Resultados Relevantes	Validación empírica mediante modelos de jerarquía multinivel y exclusiones mutuas en clasificación de imágenes finas.

13. Tutorial sobre Métodos de Atribución Basados en Gradientes para Redes Neuronales Profundas (Nielsen et al., 2022).

La capacidad de comprender y explicar las decisiones de estos modelos no solo es importante para construir confianza con los usuarios, sino también para cumplir con regulaciones legales como el Reglamento General de Protección de Datos (GDPR) de la Unión Europea. El documento trata diferentes métodos de atribución basados en gradientes, destacando mejoras recientes en la visualización y robustez de los mapas de atribución. A continuación, se presenta un resumen de los aspectos clave tratados en el documento, incluyendo mejoras en la explicabilidad, comparación con enfoques anteriores y los impactos de estas mejoras en la interpretación de modelos.

El documento se centra en la explicabilidad de los métodos de atribución basados en gradientes. Se aborda la importancia de los mapas de atribución y su sensibilidad a las perturbaciones adversariales. (Véase Tabla 31)

Tabla 31. Método de atribución basado en Gradientes.

Aspecto	Descripción
Mejoras en la Explicabilidad	Se destacan mejoras como la reducción del ruido y la difusión visual mediante métodos como SmoothGrad y Grad-CAM. Estos métodos proporcionan una mejor visualización y comprensión de las contribuciones de cada característica a la salida del modelo .
Marco	El marco teórico se basa en la utilización de mapas de atribución para redes neuronales, explorando métodos como Gradientes, DeconvNets, Backpropagation Guiada y Grad-CAM. Se investiga la relación entre la robustez adversarial y la interpretabilidad de los mapas de atribución .

Comparación con Enfoques Anteriores	El documento compara varios métodos de atribución, destacando las limitaciones de los mapas de saliencia en términos de ruido visual y proponiendo métodos como SmoothGrad para mejorar la claridad. También se compara la robustez adversarial de las redes neuronales entrenadas con métodos tradicionales frente a los entrenados con técnicas adversariales .
Impacto de las mejoras	Las mejoras propuestas permiten una mejor alineación visual de los mapas de atribución con la percepción humana. Se observa que los modelos entrenados adversarialmente no sólo son más robustos, sino que también producen mapas de atribución más interpretables y menos ruidosos
Propuestas y Taxonomías	El documento propone una taxonomía de métodos de atribución basados en gradientes y destaca la importancia de la elección adecuada de hiperparámetros y puntos de referencia para garantizar explicaciones válidas. Se discute la necesidad de métodos que sean invariantes a la implementación y robustos frente a perturbaciones adversariales .
Resultados Relevantes	Los resultados experimentales muestran que los modelos robustos frente a ataques adversariales generan mapas de atribución que son más coherentes con las características relevantes de los datos de entrada. Además, se demuestra que la robustez adversarial contribuye a la interpretabilidad de los modelos, mejorando la alineación de los mapas de atribución con las expectativas humanas.

14.Explicando la Caja Negra: Un Enfoque Contrafactual (Singla et al. (2023)

Se presenta un método para explicar modelos de clasificación de imágenes en aplicaciones médicas. En contraste con enfoques clásicos que evalúan la importancia de características

(por ejemplo, mapas de saliencia), el marco explica cómo las características de imagen en regiones anatómicas importantes son relevantes para la decisión de clasificación. Se utilizó una Red Generativa Adversaria (GAN) para generar una serie progresiva de perturbaciones a una imagen de consulta, modificando gradualmente la decisión de clasificación original a su negación. La función de pérdida preserva detalles esenciales (como dispositivos de soporte) en las imágenes generadas (Véase Tabla 32).

Se emplearon explicaciones contrafactuales para auditar un clasificador entrenado en un conjunto de datos de radiografías de tórax con múltiples etiquetas. La evaluación clínica de las explicaciones del modelo es un desafío; por lo tanto, se propusieron métricas cuantitativas clínicamente relevantes, como la proporción cardio-torácica y el puntaje del seno costofrénico sano, para evaluar las explicaciones. Estas métricas se utilizaron para cuantificar los cambios contrafactuales entre las poblaciones con decisiones negativas y positivas para un diagnóstico según el clasificador dado.

Se llevó a cabo un experimento fundamentado en humanos con residentes de radiología diagnóstica para comparar diferentes estilos de explicaciones (sin explicación, mapa de saliencia, explicación de CycleGAN y la explicación contrafactual), evaluando diferentes aspectos: (1) comprensibilidad, (2) justificación de la decisión del clasificador, (3) calidad visual, (4) preservación de la identidad y (5) utilidad general de una explicación para los usuarios. El resultado indica que la explicación contrafactual fue el único método que mejoró significativamente la comprensión de los usuarios sobre la decisión del clasificador en comparación con el caso sin explicación. Las métricas establecieron un punto de referencia para evaluar métodos de explicación de modelos en imágenes médicas. Las explicaciones revelaron que el clasificador se basaba en características radiográficas clínicamente relevantes para sus decisiones diagnósticas, haciendo su proceso de toma de decisiones más transparente para el usuario final.

Tabla 32. Análisis del Método de Explicación Contrafactual para Modelos de Clasificación de Imágenes Médicas.

Aspecto	Descripción
Mejoras en la Explicabilidad	Desarrollo de un Método de Explicación Contrafactual para Cajas Negras en modelos de clasificación de imágenes médicas. Enfoque en explicar la relevancia de características de imágenes anatómicas críticas para decisiones de clasificación transparentes en salud. Utilización de GAN para generar perturbaciones progresivas y

	una función de pérdida para preservar detalles esenciales en las imágenes generadas.
Marco	Uso de una Red Generativa Adversaria (GAN) para crear perturbaciones graduales en imágenes de consulta y explicar decisiones de clasificación.
Comparación con Enfoques Anteriores	Contraste con métodos tradicionales como los mapas de saliencia, destacando la capacidad de explicar cómo las características radiográficas relevantes clínicamente influyen en las decisiones del clasificador.
Impacto de las mejoras	Mejora significativa en la comprensión de las decisiones del clasificador por parte de usuarios médicos, en comparación con métodos de explicación como los mapas de saliencia y explicaciones cicloGAN, según evaluaciones humanas.
Propuestas y Taxonomías	Establecimiento de métricas cuantitativas clínicamente relevantes, como la proporción cardio-torácica y la evaluación de la fosa costofrénica normal, para medir cambios contrafactuales entre poblaciones con decisiones negativas y positivas del clasificador.
Resultados Relevantes	Revelación de la dependencia del clasificador en características radiográficas clínicamente relevantes para decisiones diagnósticas, mejorando la transparencia del proceso de toma de decisiones. Evaluación establecida como punto de referencia para métodos de explicación en imágenes médicas.

15. Modelos de cuello de botella conceptual post-hoc (Yuksekgonul, Wang, & Zou, 2022).

El artículo introduce los modelos de Post-hoc Concept Bottleneck (P-CBM) como una extensión de los Concept Bottleneck Models (CBM) que busca mejorar la interpretabilidad de

los modelos de aprendizaje profundo sin sacrificar su rendimiento. Mientras que los CBMs requieren etiquetas de conceptos durante el entrenamiento, limitando su aplicabilidad práctica, los P-CBMs permiten convertir cualquier modelo preentrenado en un CBM de manera retroactiva, utilizando una biblioteca de conceptos para mejorar la interpretación del modelo.

Los P-CBMs funcionan proyectando las representaciones de entrada a un espacio de conceptos definido por vectores de activación de conceptos (CAVs), aprendidos a partir de ejemplos positivos y negativos. Esto permite examinar qué conceptos están utilizando los modelos para tomar decisiones, facilitando la intervención y la edición del modelo sin necesidad de ajuste fino en el dominio de destino. Además, se introduce el modelo híbrido HP-CBM para manejar casos donde el banco de conceptos inicial no es suficiente, utilizando un predictor residual para mejorar la precisión del modelo original.

En términos de resultados experimentales, los P-CBMs y HP-CBMs muestran una pérdida mínima de rendimiento en comparación con los modelos originales en diversas tareas de clasificación de imágenes y médicas, demostrando su utilidad y robustez. Además, se demuestra que los P-CBMs permiten ediciones simples en el espacio de conceptos para mejorar el rendimiento del modelo sin necesidad de reentrenamiento o datos adicionales del dominio de prueba.

En el análisis de los avances recientes en la explicabilidad de los modelos de machine learning, se destacan los Post-hoc Concept Bottleneck Models (P-CBMs), presentados por Yuksekgonul, Wang y Zou (2022). Estos modelos mejoran la interpretabilidad sin sacrificar su rendimiento. La Tabla 33 detalla los aspectos clave y los resultados relevantes de esta innovación.

Tabla 33.Aspectos y Resultados de los Modelos Post-hoc Concept Bottleneck (P-CBM)

Aspecto	Descripción
Mejoras en la Explicabilidad	Introduce P-CBMs como modelos que mejoran la interpretabilidad de modelos de aprendizaje profundo al permitir la interpretación de conceptos utilizados para la toma de decisiones.
Marco	Los P-CBMs se basan en la idea de proyectar representaciones de entrada en un espacio de conceptos definido por vectores de activación de conceptos (CAVs),

	aprendidos a partir de ejemplos positivos y negativos, sin necesidad de etiquetas de conceptos durante el entrenamiento inicial.
Comparación con Enfoques Anteriores	Contrasta con los CBMs tradicionales que requieren etiquetas de conceptos durante el entrenamiento, limitando su aplicabilidad práctica. Además, se compara con otros enfoques de explicabilidad que no logran combinar interpretabilidad con rendimiento comparable al modelo original.
Impacto de las Mejoras	Permite la intervención y edición del modelo en el espacio de conceptos para mejorar su rendimiento sin ajuste fino en el dominio de destino. Mejora la comprensión de qué conceptos son utilizados por el modelo, facilitando la identificación de errores y la optimización del desempeño.
Propuestas y Taxonomías	Propone la taxonomía de P-CBMs e introduce el modelo híbrido HP-CBM para manejar limitaciones en el banco de conceptos inicial.
Resultados Relevantes	Experimentos muestran que los P-CBMs y HP-CBMs mantienen una pérdida mínima de rendimiento en comparación con los modelos originales en diversas tareas, destacando su utilidad en aplicaciones de clasificación de imágenes y médicas.

16. sMRI-PatchNet: Una novedosa red de aprendizaje profundo basada en parches explicables y eficientes para el diagnóstico de la enfermedad de Alzheimer con resonancia magnética.

sMRI-PatchNet, es una red neuronal convolucional basada en parches diseñada para la extracción de características y la clasificación en el contexto del diagnóstico de la enfermedad de Alzheimer utilizando imágenes de resonancia magnética estructural (sMRI). A continuación se detalla el funcionamiento del modelo:

Selección de Parches: Antes de ingresar a sMRI-PatchNet, se realiza una selección de parches explicables los cuales son áreas específicas dentro de las imágenes sMRI que se consideran relevantes para el diagnóstico de AD.

Flattening de Parches: Una vez seleccionados, los parches se convierten en vectores planos. Cada parche seleccionado se representa como un vector $(x_p \in R^{M\times(P^3)})$, donde M es la dimensión de los parches seleccionados.

Proyección Lineal y Embedding de Posición: Cada vector de parche x_p se somete a una proyección lineal. Además, se agrega un embedding de posición aprendido para mantener la información espacial y de ubicación de los parches dentro de la imagen sMRI original.

Bloques CNN: sMRI-PatchNet incluye dos bloques principales de redes convolucionales:

- **Global Spatial Information (GSI)**: Este bloque se encarga de capturar información sobre las relaciones espaciales globales entre los diferentes parches seleccionados. Ayuda a la red a entender cómo se distribuyen los parches seleccionados en la imagen completa de sMRI.
- **Local Patch Information (LPI)**: Este bloque está diseñado para capturar características locales dentro de cada parche individualmente seleccionado. Esto es importante ya que permite identificar detalles específicos dentro de cada área del cerebro que podrían estar relacionados con AD.

Capa Clasificadora: Después de pasar por los bloques CNN, la información se procesa a través de una capa de pooling promedio para combinar las características extraídas y luego a través de una capa completamente conectada. Esta capa final se encarga de clasificar la imagen de sMRI en las categorías deseadas, que pueden incluir Alzheimer vs. grupo normal y MCI progresivo vs. MCI estable.

Avances Recientes en la Explicabilidad

1. Uso de SHAP (SHapley Additive exPlanations):
    - El método propuesto utiliza SHAP para identificar las ubicaciones de parches informativos que son relevantes para la clasificación del Alzheimer. Este enfoque permite interpretar cuáles áreas del cerebro son más importantes para el diagnóstico, lo cual mejora la transparencia del modelo.
2. Comparación con Métodos Tradicionales:
    - El estudio compara su método con varios métodos tradicionales basados en machine learning, como SVM, LDA y KNN, y encuentra que los métodos basados en deep learning, como el propuesto PatchNet, tienen un mejor rendimiento debido a su capacidad para manejar características espaciales complejas.

Impacto de las Mejoras

1. Desempeño Competitivo:
    - El método propuesto logra un rendimiento competitivo en las tareas de clasificación relacionadas con el Alzheimer, especialmente en la predicción de la conversión de MCI (Mild Cognitive Impairment), lo cual indica una mejora significativa en comparación con métodos anteriores.
2. Reducción de Complejidad Computacional:
    - Aunque se utilizan convoluciones 3D en otros métodos, el sMRI-PatchNet utiliza convoluciones 2D, lo que resulta en una menor complejidad computacional sin comprometer el rendimiento del modelo.
3. Automatización en la Selección de Regiones de Interés (ROIs):
    - A diferencia de los métodos basados en ROI que dependen de la selección manual de regiones de interés, el método propuesto automatiza la extracción de áreas importantes a partir de múltiples parches distribuidos en todo el cerebro. Esto mejora la objetividad y reproducibilidad del modelo.

Propuestas y Taxonomías

- PatchNet: Proponen una red eficiente basada en parches que selecciona automáticamente las regiones discriminativas utilizando valores de SHAP, lo que no solo mejora la interpretabilidad sino también la eficiencia computacional.
- Comparación de Complejidad: Se compara la complejidad computacional y el número de parámetros entre diferentes métodos de deep learning, destacando la eficiencia del enfoque.

Resultados Relevantes

- Resultados en Tablas: El documento presenta comparaciones detalladas del rendimiento del modelo propuesto en relación con otros modelos del estado del arte, demostrando mejoras en precisión y eficiencia (Véase tabla 34).

Tabla 34. Evaluación de Mejoras en la Explicabilidad y Desempeño del Modelo sMRI-PatchNet.

Aspecto	Descripción
Mejoras en la Explicabilidad	**Interpretación de Modelos**: Uso de métodos como SHAP (SHapley Additive exPlanations) para identificar y explicar qué características de las imágenes de resonancia magnética estructural (sMRI) son más relevantes para la clasificación del

	Alzheimer. **Reducción de la Opacidad**: Mejoras como las implementadas en sMRI-PatchNet ayudan a abrir estas cajas negras al proporcionar explicaciones claras sobre cómo se llega a una predicción o clasificación. **Automatización de la Explicabilidad**: Implementación de técnicas que automatizan la extracción y presentación de explicaciones. En el caso de sMRI-PatchNet, esto se logra mediante la selección automática de parches relevantes y la visualización de cómo contribuyen estos parches a las decisiones del modelo. **Mejora en la confiabilidad**: Al hacer que los modelos sean más interpretables, se mejora la confiabilidad de sus predicciones. Los médicos y los investigadores pueden confiar más en los resultados del modelo cuando entienden claramente el razonamiento detrás de cada predicción.
Marco	sMRI-PatchNet es una red neuronal convolucional basada en parches diseñada para el diagnóstico de la enfermedad de Alzheimer utilizando imágenes de resonancia magnética estructural (sMRI).
Comparación con Enfoques Anteriores	El método utiliza SHAP (SHapley Additive exPlanations) para identificar parches relevantes, comparándose favorablemente con métodos tradicionales como SVM, LDA y KNN en términos de rendimiento y capacidad para manejar características espaciales complejas.
Impacto de las Mejoras	**Desempeño Competitivo**: Mejora significativa en la predicción de la conversión de MCI, mostrando un rendimiento competitivo en tareas relacionadas con el Alzheimer. **Reducción de Complejidad Computacional**: Utilización de

	convoluciones 2D en lugar de 3D, lo que reduce la complejidad computacional sin comprometer el rendimiento. **Automatización en la Selección de ROIs**: Automatización en la extracción de áreas importantes a partir de múltiples parches distribuidos en todo el cerebro, mejorando la objetividad y reproducibilidad.
Propuestas y Taxonomías	Propuesta de PatchNet: Introduce una red eficiente basada en parches que selecciona automáticamente regiones discriminativas usando SHAP, mejorando la interpretabilidad y eficiencia computacional. Comparación de Complejidad: Análisis detallado de la complejidad computacional y número de parámetros comparado con otros métodos de deep learning.
Resultados Relevantes	Incluye tablas detalladas que comparan el rendimiento del modelo sMRI-PatchNet con otros modelos del estado del arte, demostrando mejoras en la precisión y la eficiencia.

17. Aprendizaje Contrastivo Supervisado para Coincidencia de Documentos en Formato Largo Interpretable.

CoLDE opera proporcionando un puntaje de similitud ponderado entre diferentes segmentos y secciones dentro y entre los documentos. Esto significa que no solo evalúa la similitud global entre documentos completos, sino que también considera cómo se relacionan diferentes partes y secciones dentro de cada documento y entre documentos diferentes.

Avances Recientes en la Explicabilidad de Modelos 'Caja Negra'

1. CoLDE: Contrastive Long Document Encoder.
    - **Descripción del Modelo**: CoLDE es un marco basado en transformadores diseñado para emparejar documentos largos de manera interpretable. Utiliza incrustaciones posicionales únicas y una capa de atención multi-cabezal junto con un marco de aprendizaje contrastivo supervisado.
    - **Mejora en la Explicabilidad**: A diferencia de los modelos anteriores que operan principalmente con documentos cortos, CoLDE aborda desafíos

específicos de documentos largos, como la heterogeneidad contextual y la medida global de similitud inadecuada. Esto se logra mediante la captura de similitudes a múltiples niveles: entre documentos completos, secciones dentro de documentos y fragmentos individuales.

- **Impacto**: CoLDE proporciona puntuaciones de similitud detalladas que mejoran la interpretabilidad al permitir comparaciones precisas y contextualmente relevantes entre documentos largos. Esto es crucial para aplicaciones donde entender la similitud precisa entre secciones de documentos extensos es fundamental, como en la investigación científica, documentos legales y patentes.

2. Comparación con Enfoques Anteriores.
   - Se comparó el modelo CoLDE con varios baselines reconocidos en la tarea de emparejamiento de documentos largos, incluyendo DSSM, ARC-I, Hierarchical Attention Networks (HAN), Siamese-BERT (S-BERT), SMITH y S-LONG.
   - Mejoras de CoLDE: CoLDE supera estas limitaciones al incorporar incrustaciones posicionales específicas y una atención chunkwise que permite capturar relaciones finas entre partes de documentos. Esto contrasta con los enfoques más simples que usan medidas de similitud global.

Impacto en Propuestas, Taxonomías y Otros Resultados Relevantes.

1. **Propuestas Nuevas**: CoLDE propone un enfoque avanzado para el emparejamiento de documentos largos que integra técnicas de aprendizaje contrastivo supervisado, mejorando tanto la precisión como la explicabilidad del modelo.
2. **Taxonomías**: Introduce una taxonomía de explicabilidad en el contexto de modelos de emparejamiento de documentos, destacando la importancia de las medidas de similitud contextuales y detalladas.
3. **Resultados Relevantes**: CoLDE demuestra mejoras significativas sobre los métodos estado del arte en términos de precisión y robustez, además de proporcionar resultados interpretables que pueden ser crucialmente utilizados en aplicaciones donde se requiere una comprensión detallada de las similitudes entre documentos largos (Véase la tabla 35).

Tabla 35.Avances en la Explicabilidad y Eficiencia de CoLDE en el Emparejamiento de Documentos Largos.

Aspecto	Descripción
Mejoras en la Explicabilidad	CoLDE introduce un enfoque de

	aprendizaje contrastivo supervisado y atención chunkwise multi-cabezal. Este enfoque permite una interpretación más detallada al capturar similitudes semánticas a diferentes niveles dentro de documentos largos, incluyendo el documento completo, secciones específicas y fragmentos individuales.
Marco	El marco de CoLDE se basa en adaptaciones clave de las redes transformer, específicamente diseñadas para manejar documentos largos con estructuras jerárquicas y variabilidad contextual. Al integrar una función de pérdida contrastiva supervisada y atención multi-cabezal chunkwise, CoLDE estructura los documentos en secciones y fragmentos, mejorando la captura de similitudes semánticas precisas y facilitando la interpretación de cómo el modelo toma decisiones en diferentes contextos documentales.
Comparación con Enfoques Anteriores	Se comparó el modelo CoLDE con varios baselines reconocidos en la tarea de emparejamiento de documentos largos, incluyendo DSSM, ARC-I, Hierarchical Attention Networks (HAN), Siamese-BERT (S-BERT), SMITH y S-LONG.
Impacto de las Mejoras	Las mejoras introducidas por CoLDE permiten una interpretación más clara y detallada de las decisiones del modelo en comparación con métodos anteriores.
Propuestas y Taxonomías	CoLDE propone la utilización de aprendizaje contrastivo supervisado y atención chunkwise multi-cabezal como un marco efectivo para mejorar la explicabilidad en modelos de documentos largos. Introduce una taxonomía de interpretación a varios niveles (documento, sección, fragmento) que puede ser aplicada para entender cómo y por qué el modelo

	toma ciertas decisiones.
Resultados Relevantes	Los resultados de CoLDE muestran un rendimiento superior en la tarea de emparejamiento de documentos largos en comparación con métodos anteriores. La capacidad del modelo para capturar variaciones contextuales dentro de los documentos y proporcionar puntuaciones de similitud finamente detalladas ha sido validada en conjuntos de datos de documentos académicos, artículos de Wikipedia y patentes de USPTO.

18.Un Marco de Aprendizaje Contrastivo Multi-expertos Centrado en Preguntas para Mejorar la Precisión y la Interpretabilidad de los Modelos de Rastreo de Conocimiento Secuencial Profundo (Jha, Rakesh, Chandrashekar, Samavedhi, & Reddy, 2023).

A continuación se analizarán las propuestas recientes, las mejoras en la explicabilidad, y se comparan con enfoques anteriores, además de discutir el impacto en taxonomías y otros resultados relevantes.

Propuestas Recientes:

Enfoque Centrado en Preguntas: El paper introduce un marco de aprendizaje contrastivo con múltiples expertos centrado en preguntas, lo que representa una nueva propuesta para mejorar la explicabilidad de los modelos de machine learning.

Aprendizaje Contrastivo: Utiliza aprendizaje contrastivo para mejorar la capacidad del modelo de discernir entre diferentes explicaciones posibles, lo que puede proporcionar explicaciones más precisas y relevantes.

Comparación con Enfoques Anteriores:

Mejoras en la Explabilidad: El paper destaca cómo su enfoque centrado en preguntas y el uso de múltiples expertos contrastan con métodos anteriores que pueden ser menos dinámicos y adaptativos en términos de generación de explicaciones.

Innovación en Técnicas: La introducción de un marco de aprendizaje contrastivo representa un avance sobre técnicas tradicionales que no explotan este tipo de aprendizaje para la explicabilidad.

Para demostrar la efectividad del método, se realiza una comparación con varios métodos existentes (véase tabla 36):

LIME (Local Interpretable Model-agnostic Explanations):

- LIME genera explicaciones locales al aproximar el modelo complejo con un modelo más simple en la vecindad de la predicción a explicar.

SHAP (SHapley Additive exPlanations):

- SHAP se basa en valores de Shapley de teoría de juegos para asignar la importancia de cada característica en la predicción, proporcionando explicaciones consistentes y aditivas.

Grad-CAM (Gradient-weighted Class Activation Mapping):

- Grad-CAM utiliza gradientes de las características finales para producir mapas de activación de clase, visualizando qué partes de la imagen influyen en la predicción del modelo.

Integrated Gradients:

- Este método atribuye la importancia de las características calculando el gradiente a lo largo del camino desde una referencia básica hasta la entrada actual, integrando estas contribuciones.

Anchors:

- Anchors proporciona reglas de decisión de alto anclaje que son suficientes para asegurar una predicción similar en la mayoría de los casos.

Counterfactual Explanations:

- Este enfoque genera ejemplos contra-factuales que modifican la entrada mínima para cambiar la predicción del modelo, ofreciendo información sobre las decisiones del modelo.

Se evalúa la comparación con los métodos mencionados para demostrar la superioridad y las mejoras del método propuesto. Los aspectos claves de esta comparación son:

Evaluación Empírica:

El paper realiza experimentos empíricos comparando el rendimiento del marco propuesto con los métodos existentes como LIME, SHAP, Grad-CAM, Integrated Gradients, Anchors y Counterfactual Explanations.

Se miden métricas de desempeño como precisión de las explicaciones, relevancia para los usuarios, robustez frente a cambios en los datos y la capacidad de generar explicaciones coherentes.

Análisis Cuantitativo y Cualitativo:

Se presentan análisis cuantitativos, donde se utilizan datasets estándar y métricas específicas para evaluar la efectividad de las explicaciones generadas por cada método.

También se realiza un análisis cualitativo que incluye estudios de caso y ejemplos ilustrativos para demostrar cómo el método propuesto proporciona explicaciones más útiles y comprensibles.

Estudios de Usuario:

Para evaluar la relevancia y utilidad de las explicaciones desde la perspectiva del usuario, se llevan a cabo estudios con usuarios que comparan las explicaciones generadas por los diferentes métodos.

Los resultados de estos estudios indican que el método propuesto es preferido por los usuarios debido a su claridad y precisión.

Robustez y Generalización:

Se evalúa la robustez del método propuesto frente a diferentes tipos de perturbaciones y cambios en los datos de entrada.

La capacidad del marco para generalizar a diferentes dominios y tipos de datos también se analiza en comparación con los métodos existentes.

Impacto en Taxonomías y Resultados Relevantes:

Nuevo Marco Taxonómico: El enfoque presentado puede influir en la taxonomía de métodos explicativos al introducir categorías basadas en la utilización de múltiples expertos y la generación de explicaciones centradas en preguntas.

Resultados Relevantes: Los resultados obtenidos con este nuevo enfoque pueden demostrar mejoras en la precisión y relevancia de las explicaciones proporcionadas por los modelos, lo cual es un resultado relevante en el campo de la explicabilidad.

Tabla 36.Análisis de Mejora y Comparación del Método Q-MCKT para Explicabilidad de Modelos de Machine Learning.

Aspecto	Descripción
Mejoras en la Explicabilidad	Modelo Centrado en Preguntas: Descripción: El marco Q-MCKT se enfoca en modelar el estado de adquisición de conocimiento de los estudiantes a nivel de pregunta y concepto. Propósito: Capturar la variabilidad en la adquisición de conocimiento en preguntas que comparten el mismo conjunto de componentes de conocimiento (KCs). Uso de Múltiples Expertos: Descripción: Se utiliza la técnica de mezcla de expertos para mejorar la representación del conocimiento adquirido en diferentes niveles. Propósito: Mejorar la robustez y precisión del estado de adquisición de conocimiento al utilizar múltiples modelos especializados. Aprendizaje Contrastivo Centrado en Preguntas: Descripción: Se introduce una tarea de aprendizaje contrastivo detallada centrada en las preguntas. Propósito: Mejorar las representaciones de preguntas con menor interacción y aumentar la precisión de los estados de adquisición de conocimiento correspondientes. Capa de Predicción Basada en la Teoría de Respuesta a Ítems (IRT): Descripción: Se utiliza una capa de predicción basada en la teoría de respuesta a ítems para generar resultados de predicción interpretables. Propósito: Facilitar la interpretación de los resultados de las predicciones para los educadores.

	Evaluación y Reproducibilidad: Descripción: Se evalúa el marco Q-MCKT en cuatro conjuntos de datos educativos públicos del mundo real. Propósito: Demostrar que el enfoque propuesto supera a una variedad de modelos basados en aprendizaje profundo en términos de precisión y mantenibilidad de la interpretabilidad del modelo.
Marco	Se presenta un marco llamado Q-MCKT(Question-centric Multi-experts Contrastive Learning framework for Knowledge Tracing, por sus siglas en inglés).
Comparación con Enfoques Anteriores	LIME, SHAP, Grad-CAM, Integrated Gradients, Anchors y Counterfactual Explanations.
Impacto de las Mejoras	La comparación con métodos existentes (LIME, SHAP, Grad-CAM, Integrated Gradients, Anchors, Counterfactual Explanations) demuestra la superioridad del método propuesto en términos de precisión, relevancia y claridad de las explicaciones.
Propuestas y Taxonomías	El enfoque puede influir en la **taxonomía de métodos explicativos** al introducir categorías basadas en múltiples expertos (la técnica de múltiples expertos, como se aplica en el enfoque Q-MCKT, introduce una nueva forma de organizar los métodos explicativos. En lugar de usar un único modelo para explicar todo, se utilizan varios modelos especializados (expertos) para abordar diferentes aspectos del problema. Esto puede llevar a una categoría específica en la taxonomía de métodos explicativos que se enfoca en el uso de múltiples modelos especializados) y explicaciones centradas en preguntas (el enfoque introduce una categoría centrada en la pregunta. Esto significa que las explicaciones se adaptan y personalizan en

	función de las preguntas específicas que se hacen al modelo, en lugar de proporcionar explicaciones generales. Esta perspectiva puede llevar a una nueva categoría en la taxonomía que se enfoca en cómo las explicaciones se centran en preguntas individuales en lugar de en el modelo en su totalidad).
Resultados Relevantes	Los resultados indican mejoras en la precisión y relevancia de las explicaciones proporcionadas, con evidencia empírica y estudios de usuario que destacan la claridad y precisión del método propuesto.

19. Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank (Sipos, Schäfer, Glinka, & Müller-Birn, 2023).

El artículo aborda la necesidad de mejorar la explicabilidad en la Inteligencia Artificial (IA) mediante un enfoque centrado en el usuario, denominado **Human-Centered Explainable Artificial Intelligence (HC-XAI)**.Aunque las explicaciones desarrolladas por expertos en IA se enfocan en la transparencia algorítmica, a menudo no satisfacen las necesidades de los usuarios sin experiencia en IA.

El estudio utiliza el **XAI Question Bank (XAIQB)**, una herramienta que propone preguntas que los usuarios finales podrían hacer al interactuar con sistemas de IA. Sin embargo, la aplicación práctica de XAIQB ha mostrado limitaciones y falta de cobertura de algunas necesidades de explicación.

Como resultado, se ha extendido el XAIQB con 11 nuevas preguntas y mejorado las descripciones de todas las existentes. Esta extensión busca facilitar el uso de XAIQB para investigadores y profesionales en HCI (Interacción Humano-Computadora)

Avances Recientes en la Mejora de la Explicabilidad de Modelos de Machine Learning Caja Negra (véase tabla 37).

- Enfoque Human-Centered Explainable Artificial Intelligence (HC-XAI)

Visión Socio-Técnica: La HC-XAI se centra en adaptar las explicaciones de los sistemas de IA a las necesidades de los usuarios finales, en lugar de enfocarse únicamente en la transparencia algorítmica. Este enfoque considera tanto a expertos como a usuarios no

expertos, asegurando que las explicaciones sean relevantes y comprensibles para diferentes grupos de usuarios .

- XAI Question Bank (XAIQB)

Conjunto de Preguntas: El XAIQB es una herramienta que proporciona una serie de preguntas que los usuarios finales pueden hacer al interactuar con un sistema de IA. Esto ayuda a los desarrolladores a identificar y abordar las necesidades de explicación desde la perspectiva de los usuarios .

- Extensión de XAIQB

Nuevas Preguntas y Descripciones Mejoradas: La extensión del XAIQB con 11 nuevas preguntas y descripciones más detalladas aborda las deficiencias identificadas en estudios previos y mejora la capacidad de la herramienta para capturar las necesidades de explicación en contextos específicos .

Comparación con Enfoques Anteriores

- Enfoques Anteriores

Falta de Enfoque en el Usuario Final: Los enfoques anteriores se centraban en la transparencia algorítmica y en la comprensión interna de los sistemas de IA, sin considerar adecuadamente las necesidades de los usuarios no expertos. Esto resulta en explicaciones que no son efectivas o comprensibles para los usuarios finales.

- Mejoras Recientes

Adaptación a Necesidades Específicas: La HC-XAI y la extensión del XAIQB representan un avance significativo al integrar las necesidades de los usuarios finales en el proceso de explicación. Estas mejoras permiten que las explicaciones sean más relevantes y útiles para una amplia gama de usuarios, en lugar de centrarse únicamente en la complejidad algorítmica.

Impacto de las Mejoras

- Propuestas

Herramientas Mejoradas: La extensión del XAIQB y la introducción de nuevas preguntas permiten a los investigadores y diseñadores adaptar las explicaciones a las necesidades específicas de los usuarios finales. Esto facilita la identificación y abordaje de las necesidades de explicación en diferentes contextos.

- Taxonomías

Nueva Taxonomía de Preguntas: La adición de nuevas preguntas al XAIQB contribuye a una taxonomía más completa de necesidades de explicación, permitiendo una clasificación más detallada y específica de las necesidades de los usuarios .

Otros Resultados Relevantes

Mejora en la Aplicabilidad Práctica: La extensión y mejora de XAIQB facilitan su uso práctico en diversos contextos, como lo demuestra su aplicación en el estudio con historiadores del arte. Esto sugiere que la herramienta puede ser útil en una variedad de dominios y para diferentes tipos de usuarios.

Los avances recientes en la mejora de la explicabilidad de modelos de machine learning se centran en adaptar las explicaciones a las necesidades de los usuarios finales mediante enfoques como HC-XAI y la extensión de XAIQB. Estos avances superan las limitaciones de enfoques anteriores al proporcionar explicaciones más relevantes y comprensibles para una variedad de usuarios. La extensión del XAIQB con nuevas preguntas y descripciones más detalladas mejora su capacidad para capturar y abordar las necesidades de explicación en contextos específicos, lo que tiene un impacto significativo en la practicidad y aplicabilidad de las explicaciones de IA.

Tabla 37.Avances en la Identificación y Mejora de Explicabilidad para Usuarios Finales.

Aspecto	Descripción
Mejoras en la explicabilidad	La adopción del enfoque Human-Centered Explainable Artificial Intelligence (HC-XAI) busca adaptar las explicaciones a las necesidades de los usuarios finales, en lugar de enfocarse únicamente en la transparencia algorítmica. La extensión del XAI Question Bank (XAIQB) con 11 nuevas preguntas mejora la capacidad para capturar y abordar las necesidades de explicación en contextos específicos.
Marco	HC-XAI se centra en mejorar la interpretación de las decisiones de los sistemas de IA para los usuarios finales. El XAIQB proporciona un conjunto de preguntas diseñado para identificar y

	abordar las necesidades de explicación desde la perspectiva del usuario.
Comparación con Enfoques Anteriores	Los enfoques anteriores se enfocan en la transparencia algorítmica y la comprensión interna de los sistemas de IA, sin considerar adecuadamente las necesidades de los usuarios no expertos. En contraste, HC-XAI y la extensión del XAIQB abordan estas limitaciones al proporcionar explicaciones más accesibles y relevantes para una variedad de usuarios.
Impacto de las Mejoras	La integración de HC-XAI y la mejora de XAIQB resultan en explicaciones más relevantes y comprensibles para una gama más amplia de usuarios. Esto mejora la utilidad práctica y la aplicabilidad del XAIQB en diversos contextos, y facilita una mejor comprensión de las decisiones de IA por parte de los usuarios finales.
Propuestas y Taxonomías	La extensión del XAIQB con nuevas preguntas y descripciones mejoradas contribuye a una taxonomía más completa de necesidades de explicación, permitiendo una clasificación más detallada de las preguntas que los usuarios pueden tener. Esto ofrece una base más sólida para adaptar las explicaciones a las necesidades específicas de los usuarios.
Resultados Relevantes	La herramienta XAIQB mejorada demuestra una mayor aplicabilidad en diversos contextos y para diferentes tipos de usuarios, como lo evidenció su uso en el análisis con historiadores del arte. La extensión y mejora de XAIQB permiten abordar deficiencias anteriores y proporcionan una mejor base para la identificación de necesidades de explicación en futuros estudios.

20. Approaching Explainable Artificial Intelligence Methods in the Diagnosis of Iron Deficiency Anemia Using Blood Parameters (Ponnusamy, B S, & Sampathila, 2023).

Flujo de Trabajo en el Diagnóstico de Anemia por Deficiencia de Hierro.

1. **Subida del Conjunto de Datos**: El primer paso en el proceso es cargar el conjunto de datos en el sistema.
2. **Relleno de Datos Faltantes**: Después de cargar los datos, se deben tratar los valores faltantes. Esto implica completar o estimar los datos que están ausentes en el conjunto.
3. **Balanceo de Datos**: Este paso se realiza para asegurar que el conjunto de datos esté equilibrado, es decir, que las diferentes clases o categorías en los datos tengan una representación equitativa. Esto es importante para evitar sesgos en el modelo de machine learning (ML).
4. **Selección de Características**: Una vez que los datos están completos y equilibrados, se procede a seleccionar las características más relevantes que se usarán para entrenar el modelo. Este proceso busca reducir la dimensionalidad del conjunto de datos y mejorar la eficiencia del modelo.
5. **Entrenamiento de Modelos de ML**: Con las características seleccionadas, se entrenan los modelos de machine learning. Este paso implica usar el conjunto de datos para enseñar al modelo a hacer predicciones o clasificaciones.
6. **Evaluación**: Después de entrenar el modelo, se evalúa su desempeño usando métricas adecuadas. Esto ayuda a entender cómo de bien está funcionando el modelo en la tarea que se le ha asignado.
7. **Explicación con XAI (Explainable AI)**: Finalmente, se aplican técnicas de inteligencia artificial explicable para interpretar y entender cómo el modelo toma decisiones. Esto es crucial para obtener confianza en el modelo y comprender los factores que influyen en sus predicciones.

Avances Recientes en la Explicabilidad:

- **SHAP (SHapley Additive exPlanations)**: Proporciona valores de Shapley para cada característica, lo que permite una interpretación consistente y global de los modelos. SHAP es capaz de explicar cualquier modelo de machine learning, ofreciendo tanto explicaciones locales como globales.
- **LIME (Local Interpretable Model-agnostic Explanations)**: Este método genera modelos locales simples alrededor de cada predicción para explicar modelos complejos, facilitando la comprensión de por qué un modelo hace una predicción específica.

- **Decision Trees and Rule-Based Systems**: Métodos que generan reglas comprensibles para los humanos a partir de datos, ayudando a desglosar las decisiones de modelos complejos en pasos lógicos y secuenciales.
- **Gradient-weighted Class Activation Mapping (Grad-CAM)**: Utilizado especialmente en redes neuronales convolucionales (CNNs) para generar mapas de calor que muestran las áreas de las imágenes que más contribuyen a la predicción del modelo.
- **Layer-wise Relevance Propagation (LRP)**: Atribuye la predicción de una red neuronal a sus características de entrada, proporcionando una descomposición de la relevancia.

2. Comparación con Enfoques Anteriores:

- **Enfoques Anteriores**: Los modelos de machine learning tradicionales, como las redes neuronales profundas (DNNs), son considerados 'caja negra' debido a su falta de interpretabilidad. Estos modelos proporcionan predicciones precisas pero sin explicaciones claras, lo que genera desconfianza y reticencia en su adopción, especialmente en áreas críticas como la medicina y la justicia.
- **Nuevos Enfoques**: Las técnicas recientes, como SHAP y LIME, han mejorado significativamente la capacidad de interpretar y explicar los modelos 'caja negra'. Estos enfoques permiten desglosar las predicciones en componentes comprensibles, proporcionando una visión clara de cómo y por qué se toman ciertas decisiones.

3. Impacto de las Mejoras:

- **Propuestas y Taxonomías**: Estas mejoras han llevado a la creación de nuevas taxonomías y frameworks para evaluar la explicabilidad de los modelos. Se han propuesto métodos para categorizar y medir la interpretabilidad, facilitando una mejor comprensión y evaluación de diferentes enfoques de XAI.
- **Resultados Relevantes**: Los avances en explicabilidad han tenido un impacto positivo en diversos campos. En medicina, por ejemplo, permiten a los médicos entender las predicciones de los modelos, lo que puede mejorar el diagnóstico y tratamiento de enfermedades. En justicia, ayudan a garantizar decisiones más justas y transparentes.
- **Confianza y Adoptabilidad**: Mejorar la explicabilidad de los modelos ha aumentado la confianza de los usuarios en estos sistemas, promoviendo una mayor adopción en diferentes sectores. La capacidad de explicar y justificar las decisiones de los modelos es crucial para su aceptación y uso ético.

La siguiente tabla (véase 38) resume las mejoras, comparaciones y resultados relevantes en la explicación de modelos de ML para el diagnóstico de anemia

Tabla 38. Comparación y Evaluación de Técnicas de Explicabilidad en el Diagnóstico de Anemia por Deficiencia de Hierro

Aspecto	Descripción
Mejoras en la explicabilidad	Implementación de técnicas XAI (SHAP) para entender la importancia de atributos en la detección de anemia. - Uso de Beeswarm plot para visualizar el impacto de las características. - Integración de un modelo de apilamiento (stacking) para mejorar la interpretación y precisión.
Marco	- Estudio del impacto de los parámetros sanguíneos en el diagnóstico de anemia utilizando métodos de ML. - Se utiliza un conjunto de datos del CBC para construir y evaluar modelos predictivos. - Aplicación de técnicas de oversampling para balancear el conjunto de datos.
Comparación con Enfoques Anteriores	- Comparación con métodos anteriores como redes neuronales, SVM, árboles de decisión y KNN. - Mejora en precisión y capacidad de clasificación en comparación con modelos tradicionales. - Implementación de modelos más avanzados y técnicas de apilamiento para superar limitaciones anteriores.
Impacto de las Mejoras	- Mejora en la precisión de diagnóstico de anemia con una precisión de hasta 100%. - Reducción del tiempo de diagnóstico y minimización de errores humanos. - Aumento de la transparencia y la confianza en los modelos de ML mediante técnicas XAI
Propuestas y Taxonomías	- Propuesta de un enfoque combinado

	utilizando XAI y modelos de apilamiento para la clasificación de anemia. - Implementación de técnicas de oversampling y selección de características para mejorar la precisión. - Uso de SHAP para interpretar el impacto de características específicas.
Resultados Relevantes	- Modelos de ML (Logistic Regression, Random Forest, SVM, KNN) muestran un rango de precisión del 80-100%. - Random Forest demuestra la mejor precisión y especificidad. - SHAP y Beeswarm plots proporcionan interpretaciones claras de la importancia de los atributos en la detección de anemia

21. IA Explicable para Medicina mediante el Intérprete de Código de ChatGPT (Kitamura, Irvan, & Yamaguchi, 2023).

En aplicaciones médicas, donde la precisión y la claridad son críticas, es fundamental que los algoritmos no sólo proporcionen respuestas correctas, sino que también sean capaces de explicar cómo y por qué se llega a dichas respuestas.

Para abordar esta necesidad, se ha propuesto un sistema de evaluación llamado Criterios de Presentación de Algoritmos Médicos (MAPC). Este sistema descompone el proceso de decisión de los algoritmos médicos en cinco factores clave, permitiendo una evaluación detallada de la explicabilidad de las respuestas generadas por ChatGPT. A través de este sistema, es posible medir si las decisiones del modelo son comprensibles en términos de lógica, aplicabilidad, selección de datos, ejecución y precisión de los resultados.

El paper analiza cómo se puede aplicar el MAPC tanto a las respuestas textuales (Text Base Prompt, TBP) como a las respuestas basadas en código (Code Base Prompt, CBP) del intérprete de código de ChatGPT. Al comparar estos dos enfoques, se demuestra que el CBP ofrece una mayor transparencia y verificabilidad, lo cual es importante para aplicaciones médicas donde cada paso del proceso de decisión debe ser claro y justificable.

A continuación, se detallan los factores de explicabilidad del MAPC y se evalúa su aplicabilidad en TBP y CBP, subrayando las ventajas del enfoque basado en código para mejorar la explicabilidad en la inteligencia artificial aplicada a la medicina.

Sistema de Criterios de Presentación de Algoritmos Médicos (MAPC).

El sistema denominado Criterios de Presentación de Algoritmos Médicos (MAPC), mide si las decisiones de ChatGPT son comprensibles en cinco factores (F1 a F5):

F1: Comprender la lógica del algoritmo médico.

F2: Evaluar la aplicabilidad del algoritmo médico al texto médico.

F3: Identificar los datos de entrada adecuados desde el texto médico para el algoritmo.

F4: Asegurar la correcta ejecución del algoritmo médico.

F5: Obtener la respuesta correcta.

El MAPC utiliza estos factores para evaluar la explicabilidad de las decisiones de ChatGPT.

Evaluación del MAPC en Respuestas de TBP y CBP

TBP (Text Base Prompt):

F5: Se puede verificar la precisión mediante la respuesta textual de ChatGPT si se sigue un formato específico.

F2: Aunque se puede inferir aplicabilidad, no se puede confirmar si ChatGPT aplicó realmente el algoritmo en su proceso de pensamiento.

F1, F3 y F4: Pueden mencionarse en la respuesta textual, pero no hay certeza de que hayan sido utilizados en la toma de decisiones.

CBP (Code Base Prompt):

F1: La lógica del algoritmo se puede verificar explícitamente mediante el código Python.

F2: La ejecución del código permite confirmar la aplicabilidad.

F3: Los valores de entrada mostrados permiten verificar la selección adecuada de datos.

F4: La correcta ejecución del algoritmo se puede observar mediante los resultados del código.

F5: Igual que en TBP, se puede verificar mediante la respuesta textual de ChatGPT.

El CBP permite una verificación más detallada y transparente, ya que se puede ejecutar el código Python para confirmar cada uno de los factores del MAPC.

Avances en la Explicabilidad de Modelos de Machine Learning:

Propuesta del CBP (Code Base Prompt): El paper introduce el **CBP** como un nuevo enfoque para mejorar la explicabilidad de los modelos de lenguaje como ChatGPT. Este método se diferencia de los enfoques anteriores, como el **TBP** (Text Base Prompt), al convertir los algoritmos médicos en código Python ejecutable. Esto permite una mayor transparencia y comprensión de cómo se toman las decisiones en el contexto médico, representando un avance en la explicabilidad de modelos que tradicionalmente son opacos.

Comparación con Enfoques Anteriores:

Text Base Prompt (TBP): El paper compara el CBP con el TBP, demostrando que el CBP supera al TBP en términos de explicabilidad. Mientras que el TBP solo proporciona respuestas basadas en texto que no permiten una verificación clara del proceso de decisión, el CBP permite la ejecución de código Python, lo que facilita la verificación de la lógica y el proceso de decisión del modelo.

Propuestas y Taxonomías:

MAPC (Medical Algorithm Presentation Criteria): Se introduce el **MAPC** como un nuevo sistema de evaluación para medir la explicabilidad en el contexto médico. Este sistema descompone el proceso de comprensión humana en cinco factores (F1 a F5), proporcionando una taxonomía específica para evaluar la calidad de la explicación en modelos como ChatGPT cuando se aplican a tareas médicas.

Impacto en Resultados Relevantes:

Resultados Experimentales: Los resultados experimentales muestran que el CBP permite verificar todos los factores del MAPC, mientras que el TBP solo cumple con uno de ellos. Esto indica que el CBP ofrece una mejora significativa en términos de explicabilidad y permite una evaluación más objetiva y transparente del modelo.

Contribuciones del Paper:

Nuevas Metodologías: El paper contribuye al campo de XAI al presentar un nuevo enfoque (CBP) y un nuevo sistema de evaluación (MAPC), ofreciendo soluciones concretas para mejorar la interpretabilidad de modelos de machine learning en aplicaciones médicas.)Véase Para resumir estos avances y compararlos con enfoques anteriores, se presenta la tabla 39.

Tabla 39. Resumen de Atributos y Contribuciones del Paper "XAI for Medicine by ChatGPT Code Interpreter.

Aspecto	Descripción
Mejoras en la explicabilidad	Introducción del CBP (Code Base Prompt) que convierte algoritmos médicos en código Python ejecutable, permitiendo una mayor transparencia y comprensión.
Marco	El CBP y el MAPC (Medical Algorithm Presentation Criteria) forman el marco metodológico para evaluar y mejorar la explicabilidad de modelos de ML en medicina.
Comparación con Enfoques Anteriores	Comparación entre CBP y TBP (Text Base Prompt), destacando que el CBP permite una verificación clara y objetiva del proceso de decisión del modelo.
Impacto de las Mejoras	El CBP supera al TBP al permitir verificar todos los factores del MAPC, mejorando significativamente la explicabilidad y la transparencia del modelo.
Propuestas y Taxonomías	Introducción del MAPC como un sistema de evaluación que descompone la comprensión humana en cinco factores (F1 a F5) para evaluar la calidad de la explicación.
Resultados Relevantes	Los resultados experimentales muestran que el CBP permite cumplir con todos los factores del MAPC, a diferencia del TBP, que solo cumple con uno.

22.Interpretación de Modelos de Machine Learning de Caja Negra para Conjuntos de Datos de Alta Dimensionalidad (Karim, Rahman, & Tareq, 2023)

A continuación se detalla el flujo de trabajo del enfoque propuesto:

Entrada de Datos: El primer paso en el flujo de trabajo generalmente involucra la recopilación o entrada de datos. En el contexto de la inteligencia artificial o el aprendizaje automático, esto podría ser un conjunto de datos de imágenes, texto, o cualquier otro tipo de datos relevante.

1. **Preprocesamiento**: Los datos a menudo requieren algún tipo de preprocesamiento antes de ser utilizados. Esto puede incluir normalización, limpieza de datos, reducción de dimensionalidad, etc.
2. **Modelo**: El siguiente paso generalmente es la construcción o aplicación de un modelo. Esto puede ser un modelo de aprendizaje profundo, un algoritmo de clasificación, o cualquier otro tipo de modelo según el enfoque propuesto.
3. **Entrenamiento**: En este paso, el modelo es entrenado utilizando los datos preprocesados. Esto implica ajustar los parámetros del modelo para minimizar la función de pérdida.
4. **Evaluación**: Después de entrenar el modelo, se evalúa su desempeño utilizando un conjunto de datos de prueba. Esto puede involucrar métricas como precisión, recall, F1-score, entre otras.
5. **Explicación**: En el contexto de XAI (Explicabilidad de la Inteligencia Artificial), puede haber una fase en la que se genera una explicación sobre cómo el modelo toma decisiones. Esto puede incluir visualizaciones, explicaciones basadas en reglas, o métodos de atribución.
6. **Resultados**: Finalmente, se presentan los resultados obtenidos y se comparan con los resultados de métodos anteriores o baselines.

El paper brinda un enfoque para mejorar la interpretabilidad de modelos complejos entrenados en conjuntos de datos de alta dimensionalidad. Propone un método que combina técnicas de sondeo y perturbación para identificar características claves a nivel global y un modelo sustituto interpretable que aproxima el comportamiento del modelo de caja negra. Esta estrategia permite no solo obtener explicaciones a nivel global, sino también derivar reglas de decisión y contrafactuales para explicaciones locales. Se demuestran mejoras en la interpretabilidad en comparación con métodos tradicionales, como los basados en SHAP y los modelos tabulares como TabNet y XGBoost.

Enfoques para Mejorar la Explicabilidad:

Modelos Sustitutos Interpretables:. El enfoque consiste en entrenar un modelo de caja negra en el espacio completo de características y luego usar técnicas de sondeo y perturbación para identificar las características más importantes (global explainability). Este método ofrece explicaciones tanto globales como locales mediante la derivación de reglas de decisión y contrafactuales.

Técnicas de Compresión y Simplificación: Reducen la complejidad y el tamaño de los modelos, facilitando su despliegue en dispositivos con recursos limitados.

Transferencia de Conocimiento y Distilación de Modelos: En lugar de simplemente entrenar un modelo sustituto basado en las predicciones del modelo negro, la transferencia de conocimiento se basa en la identificación de las características más importantes (top-k) y la optimización de una función objetivo. Este proceso transfiere el conocimiento aprendido para crear un modelo más simple y ligero, mejorando la interpretabilidad y la eficiencia.

Comparación con Enfoques Anteriores:

Métodos Basados en SHAP y Tabulares: Aunque técnicas como SHAP proporcionan interpretaciones detalladas, pueden ser computacionalmente costosas y no siempre capturan la totalidad del modelo. Los modelos tabulares como TabNet y XGBoost también se utilizan para interpretabilidad, pero el enfoque de modelos sustitutos y técnicas de compresión presentan una alternativa eficiente al abordar tanto la interpretabilidad como la eficiencia.

Métodos Basados en Gradientes y Activaciones: Los métodos tradicionales basados en gradientes, como Grad-CAM, ofrecen visibilidad en la activación de características específicas pero a menudo son limitados en términos de proporcionar explicaciones globales y comprehensivas.

Impacto de las Mejoras:

Propuestas y Taxonomías: Integración de técnicas como modelos sustitutos interpretables y métodos de compresión.

Resultados Relevantes: El uso de modelos sustitutos y técnicas de compresión no solo facilita una mejor comprensión de los modelos de caja negra, sino que también aborda problemas prácticos relacionados con el despliegue y la inferencia en aplicaciones del mundo real (Véase la tabla 40).

Tabla 40. Resumen de Avances Recientes en la Explicabilidad de Modelos de Machine Learning para Conjuntos de Datos de Alta Dimensionalidad

Aspecto	Descripción
Mejoras en la explicabilidad	**Modelos Sustitutos Interpretables:** Uso de modelos sustitutos para aproximar el comportamiento de modelos de caja negra, ofreciendo explicaciones globales y locales mediante reglas de decisión y contrafactuales. **Técnicas de Compresión:** Cuantización y poda para reducir la complejidad y el

	tamaño de los modelos, mejorando la manejabilidad y eficiencia. **Transferencia de Conocimiento:** Optimización de funciones objetivo basadas en características importantes para simplificar y mejorar la interpretabilidad de modelos complejos.
Marco	Se integran enfoques como la creación de modelos sustitutos y técnicas de compresión para abordar tanto la interpretación como la eficiencia, facilitando su despliegue en entornos con recursos limitados y mejorando la comprensión de los modelos de caja negra.
Comparación con Enfoques Anteriores	**Métodos Basados en SHAP y Tabulares:** Aunque SHAP proporciona explicaciones detalladas, es computacionalmente costoso y no siempre abarca toda la complejidad del modelo de caja negra. Los modelos tabulares como TabNet y XGBoost también son interpretables, pero los enfoques actuales con modelos sustitutos y técnicas de compresión ofrecen una alternativa más eficiente y comprensible. **Métodos Basados en Gradientes y Activaciones:** Los métodos tradicionales basados en gradientes y activaciones son útiles pero limitados en términos de explicación global y comprehensiva.
Impacto de las Mejoras	**Eficiencia Operativa**: La reducción en el tamaño y la complejidad de los modelos mediante técnicas de compresión.. **Accesibilidad de Explicaciones**:explicaciones más accesibles y comprensibles para los usuarios finales, facilitando la validación y la confianza en las decisiones del modelo. **Validación y Evaluación**: Las técnicas de surrogación y transferencia de conocimiento proporcionan una forma más efectiva de validar y evaluar los modelos de caja negra, asegurando que las explicaciones derivadas

	sean precisas y útiles.
Propuestas y Taxonomías	Los avances recientes han llevado a nuevas taxonomías y enfoques en la explicabilidad, integrando técnicas innovadoras y ampliando el marco conceptual existente.
Resultados Relevantes	Las mejoras en la interpretabilidad y eficiencia han demostrado ser significativas, facilitando la comprensión de modelos complejos y abordando problemas prácticos relacionados con el despliegue y la inferencia en aplicaciones del mundo real.

23. xAI: Un Modelo de IA Explicable para el Diagnóstico de la EPOC a partir de Imágenes de Rayos X de Tórax (Ikechukwu y Murali (2023).

El modelo propuesto combina los modelos ResNet50 y Xception mediante un método de aprendizaje por transferencia. La estructura del modelo se organiza en tres fases: (i) Pre-procesamiento de Imágenes, (ii) Aplicación del Aprendizaje por Transferencia, y (iii) Explicación del Modelo utilizando Mapeo de Activación de Clases Gradiente (Grad-CAM) y Explicaciones Aditivas de Shapley (SHAP).

Avances Recientes en la Mejora de la Explicabilidad:

- **Aplicación de Grad-CAM y SHAP:**
  La combinación de Grad-CAM (Gradient-weighted Class Activation Mapping) y SHAP (SHapley Additive exPlanations) proporciona una doble capa de explicabilidad en el diagnóstico de EPOC a partir de imágenes de radiografía de tórax (CXR).
    - **Grad-CAM** genera mapas de calor que destacan las regiones de la imagen que más contribuyen a la decisión del modelo, facilitando la visualización de qué áreas son relevantes para la predicción.
    - **SHAP** descompone la predicción del modelo en contribuciones aditivas de cada característica, permitiendo una interpretación detallada de cómo cada característica influye en la decisión del modelo.
- **Transfer Learning con Xception y ResNet50:**
  El uso del aprendizaje por transferencia con modelos preentrenados como Xception y ResNet50 mejora la capacidad del modelo para generalizar y proporcionar explicaciones más precisas. La afinación fina del modelo Xception logró un recall del

98.2%, superior al del modelo ResNet50, mostrando una mejora significativa en la precisión y explicabilidad del modelo.

Comparación con Enfoques Anteriores:

- **Métodos Tradicionales:** Las pruebas de espirometría, aunque definitivas para el diagnóstico de EPOC, no ofrecen capacidad explicativa y tienen una accesibilidad limitada en regiones con pocos recursos.
- **Enfoques Anteriores con Machine Learning:** Los modelos anteriores para el diagnóstico mediante imágenes no incluían técnicas avanzadas de explicabilidad como Grad-CAM y SHAP, lo que limitaba la confianza y comprensión de las predicciones del modelo por parte de los médicos.

Impacto de las Mejoras:

- **Propuestas y Taxonomías:** La combinación de Grad-CAM y SHAP proporciona una explicación más completa, facilitando la comprensión y confianza de los profesionales de la salud en las predicciones del modelo. Estas técnicas se han integrado en taxonomías recientes que subrayan su importancia para mejorar la transparencia y confianza en los modelos de IA.
- **Resultados Relevantes:** El estudio demuestra que el uso de Grad-CAM y SHAP en el diagnóstico de EPOC mediante CXR puede facilitar la detección temprana de la enfermedad, especialmente en áreas con limitaciones en el acceso a espirometría. La mejor explicabilidad también puede fomentar una mayor adopción de modelos de IA en la práctica clínica, permitiendo a los médicos entender y verificar las razones detrás de las predicciones del modelo.

Resumen de Técnicas Utilizadas para la Explicabilidad:

- **Grad-CAM (Gradient-weighted Class Activation Mapping):** Genera mapas de calor para resaltar las áreas de las imágenes CXR que influyen en las predicciones del modelo. Esto proporciona una forma intuitiva de entender cómo el modelo llega a sus decisiones.
- **SHAP (SHapley Additive exPlanations):** Descompone la predicción del modelo en contribuciones aditivas de cada característica, ofreciendo una explicación detallada y cuantitativa del impacto de cada característica en la predicción del modelo.

Estas técnicas combinadas mejoran significativamente la transparencia y la interpretabilidad del modelo de IA para el diagnóstico de EPOC utilizando imágenes de radiografía de tórax.(Véase tabla 41)

Tabla 41. Análisis de Explicabilidad y Comparación con Enfoques Anteriores en Diagnóstico de EPOC.

Aspecto	Descripción
Mejoras en la explicabilidad	El modelo utiliza Grad-CAM y SHAP para mejorar la interpretabilidad. Grad-CAM genera mapas de calor que destacan regiones relevantes en las imágenes CXR, mientras que SHAP proporciona una descomposición aditiva de las contribuciones de las características a la predicción del modelo.
Marco	Implementación de aprendizaje por transferencia con modelos preentrenados Xception y ResNet50, mejorando la capacidad de generalización y precisión del modelo en el diagnóstico de EPOC.
Comparación con Enfoques Anteriores	Los métodos tradicionales, como la espirometría, no ofrecen explicabilidad y tienen accesibilidad limitada. Modelos previos de machine learning no integraban técnicas avanzadas de explicabilidad como Grad-CAM y SHAP, limitando la comprensión de las predicciones.
Impacto de las Mejoras	La combinación de Grad-CAM y SHAP aumenta la confianza en las predicciones del modelo, facilita la detección temprana de EPOC, y mejora la adopción en la práctica clínica, especialmente en áreas con recursos limitados.
Propuestas y Taxonomías	La combinación de Grad-CAM y SHAP es destacada en taxonomías recientes de explicabilidad, subrayando la importancia de técnicas combinadas para una mayor transparencia en modelos médicos.
Resultados Relevantes	El uso de Grad-CAM y SHAP ha mejorado significativamente la precisión y la transparencia en el diagnóstico de EPOC, facilitando su detección temprana y la

	aceptación del modelo en la práctica clínica.

24. Inteligencia Artificial Explicable e Imagenología Cardíaca: Hacia Modelos Más Interpretables (Salih et al., 2023)

La imagenología moderna utiliza técnicas no invasivas como la resonancia magnética, la tomografía computarizada y la ecografía para evaluar la estructura y función de sistemas biológicos. Con el aumento de datos generados por estas técnicas, la inteligencia artificial (IA) ha emergido como una herramienta clave para mejorar la adquisición de imágenes, segmentación, extracción de características, diagnóstico y apoyo en la toma de decisiones (Zahid et al., 2021). Aunque el aprendizaje automático (ML) y el aprendizaje profundo (DL) han demostrado ser efectivos en estas tareas, la falta de transparencia en los modelos de DL puede limitar su adopción en entornos clínicos críticos (Choi et al., 2020).

La inteligencia artificial explicable (XAI) busca abordar esta limitación proporcionando interpretaciones claras de los modelos "caja negra" sin sacrificar el rendimiento (Keany et al., 2019). Salih et al. (2023) destacan varias metodologías de XAI relevantes, como Grad-CAM, SHAP, LIME y SmoothGrad, y discuten cómo estas técnicas pueden mejorar la interpretabilidad en el contexto de la imagenología cardíaca.

Metodologías Específicas:

- **Grad-CAM:** Utiliza gradientes de la clase objetivo para generar un mapa de calor que destaca áreas importantes en una imagen que contribuyen a la predicción.
- **SHAP:** Proporciona explicaciones locales consistentes usando valores de Shapley, permitiendo una comprensión más profunda de cómo cada característica influye en la predicción.
- **LIME:** Genera modelos interpretables localmente alrededor de cada predicción para explicar las decisiones de modelos complejos de manera más comprensible.
- **SmoothGrad:** Mejora las visualizaciones de gradientes al reducir el ruido mediante la promediación de múltiples copias ruidosas de la entrada.

Comparación con Enfoques Anteriores:

- **Modelos Clásicos (ML):** Métodos como la regresión lineal y los árboles de decisión ofrecen interpretabilidad pero con un rendimiento inferior en comparación con los modelos de DL en tareas complejas.
- **Modelos DL:** A pesar de su mayor rendimiento, los modelos de DL son menos interpretables. Los avances en XAI buscan reducir esta brecha.

Impacto y Propuestas:

- **Aplicaciones Clínicas:** La integración de XAI en la imagenología cardíaca ha permitido aplicaciones prácticas, proporcionando explicaciones visuales que mejoran la confianza de los médicos en las predicciones.
- **Propuestas de Investigación:** Se sugiere la integración continua de herramientas XAI en el flujo de trabajo clínico para superar la barrera de la "caja negra" y mejorar la evaluación de sesgos y problemas de rendimiento en modelos de IA.

La tabla 42 a continuación resume las mejoras en la explicabilidad, el marco conceptual, y la comparación con enfoques anteriores, además de destacar el impacto de estas mejoras, propuestas y taxonomías, así como los resultados relevantes en el campo.

Tabla 42. Resumen de Mejoras en la Explicabilidad y su Impacto en la Imagenología Cardíaca.

Aspecto	Descripción
Mejoras en la explicabilidad	- **Grad-CAM**: Mapa de calor basado en gradientes que destaca áreas importantes de una imagen. - **SHAP**: Explicaciones locales y consistentes usando valores de Shapley. - **LIME**: Modelos interpretables locales para explicar decisiones. - **SmoothGrad**: Mejora visualizaciones de gradientes al promediar entradas ruidosas.
Marco	- **Modelos de Caja Negra**: Modelos de deep learning (DL) son complejos y opacos, lo que dificulta la confianza. - **XAI**: Busca mejorar la interpretación de estos modelos manteniendo un alto rendimiento.
Comparación con Enfoques Anteriores	- **Modelos Clásicos (ML)**: Más interpretables (e.g., regresión lineal, árboles de decisión) pero con menor rendimiento en tareas complejas. - **Modelos DL**: Mayor rendimiento pero menos interpretables; XAI busca cerrar esta brecha.

Impacto de las Mejoras	- **Aplicaciones Clínicas**: Mejoras en XAI han permitido su uso en la imagenología cardíaca, ayudando a los médicos a entender y confiar en las predicciones. - **Confianza y Adopción**: Explicaciones claras aumentan la confianza en la IA. - **Propuestas de Investigación**: Integrar más herramientas de XAI en el flujo de trabajo clínico.
Propuestas y Taxonomías	- **Propuestas**: Integrar XAI en aplicaciones clínicas y propuestas para futuras investigaciones. - **Taxonomías**: Clasificación de métodos XAI según aplicabilidad (agnósticos vs. específicos), tipo de explicación (locales vs. globales), tipo de modelo (post-hoc vs. ante-hoc).
Resultados Relevantes	**Aplicaciones Clínicas**: Las mejoras en XAI han facilitado la implementación de herramientas de IA en la imagenología cardíaca, ofreciendo explicaciones visuales que ayudan a los médicos a interpretar las predicciones de los modelos de IA de manera más efectiva. **Confianza y Adopción**: La capacidad de proporcionar explicaciones claras y comprensibles ha aumentado la confianza de los usuarios finales en los modelos de IA. **Integración en la Práctica Clínica**: La incorporación de XAI en el flujo de trabajo clínico es importante para superar la barrera de la "caja negra", facilitando la evaluación de sesgos, modos de fallo y problemas de rendimiento en los modelos de IA.

25. Una gama de enfoques de aprendizaje automático explicables e interpretables para estudios genómicos (Conard, DenAdel, & Crawford, 2023).

El artículo aborda los avances en el uso de enfoques de aprendizaje automático explicables e interpretables en estudios genómicos. Motivados por el crecimiento exponencial de datos biológicos complejos obtenidos a través de ensayos genómicos de alto rendimiento, los autores subrayan la necesidad de metodologías estadísticas más avanzadas que los modelos

lineales tradicionales. Aunque los modelos de aprendizaje automático, como las redes neuronales, han demostrado un rendimiento superior en tareas de predicción, su naturaleza de "caja negra" limita la comprensión de cómo y por qué se obtienen dichas predicciones. Esto es especialmente problemático en biomedicina, donde la transparencia es importante para realizar pruebas de hipótesis mediante simulaciones computacionales y justificar los hallazgos del modelo para la toma de decisiones posteriores, como determinar el siguiente experimento o estrategia de tratamiento.

Para abordar esta limitación, han surgido enfoques de aprendizaje automático explicables e interpretables. Mientras que los métodos explicables buscan proporcionar una comprensión retrospectiva de lo que un modelo ha aprendido, los modelos interpretables están diseñados para ser transparentes desde su concepción. El artículo revisa el espectro de la transparencia en los modelos, desde enfoques de "caja negra" hasta modelos inherentemente interpretables, destacando avances en metodologías supervisadas y no supervisadas en genómica.

Avances Recientes en la Mejora de la Explicabilidad

Métodos Explicables Post Hoc

Buscan proporcionar una comprensión retrospectiva de lo que un modelo ha aprendido después de su entrenamiento. Estos enfoques permiten identificar la influencia de las características de entrada en las predicciones del modelo, aunque no cambian la naturaleza "caja negra" del modelo subyacente.

Ejemplos:

- **SHAP (SHapley Additive exPlanations):** Asigna valores de Shapley a cada característica, explicando la contribución de cada una a las predicciones del modelo.
- **LIME (Local Interpretable Model-agnostic Explanations):** Proporciona explicaciones locales al ajustar un modelo interpretable a las predicciones de un modelo "caja negra" en la vecindad de una instancia particular.

Modelos Intrínsecamente Interpretables (Ante-hoc)

Diseñados desde el principio para ser transparentes, estos modelos permiten una interpretación directa de sus parámetros y estructura.

Ejemplos:

- **Regresiones Lineales:** Proporcionan coeficientes claros que indican la importancia de cada característica.

- **Árboles de Decisión:** Presentan una estructura de árbol donde cada nodo representa una decisión basada en una característica, facilitando la interpretación.

Integración de Conocimiento Biológico

Incorporación de conocimiento biológico existente en el diseño de modelos interpretables.

Ejemplos:

- **Redes Parcialmente Conectadas Basadas en Anotaciones Biológicas:** Incorporan conocimiento biológico conocido en la estructura del modelo, mejorando su interpretabilidad.
- **Funciones de Pérdida Basadas en Principios Biológicos:** Utilizan funciones de pérdida que reflejan procesos biológicos, haciendo que los modelos sean más interpretables en contextos biomédicos.

Impacto de las Mejoras

Propuestas y Taxonomías:La creación de un espectro de transparencia de modelos que va desde métodos de "caja negra" a modelos completamente interpretables ha llevado a una mejor categorización y comprensión de los diferentes niveles de explicabilidad e interpretabilidad en machine learning. Esto facilita la selección de métodos adecuados según las necesidades específicas de cada aplicación.

Resultados Relevantes en Aplicaciones Biomédicas

- Identificación de Biomarcadores: Los modelos interpretables ayudan a descubrir biomarcadores importantes y entender mejor las interacciones genéticas complejas, cruciales para el desarrollo de tratamientos personalizados y la toma de decisiones clínicas basadas en modelos predictivos precisos y comprensibles.
- Validez y Confiabilidad: Los modelos interpretables permiten realizar pruebas de hipótesis estadísticas clásicas, esenciales para la validación de hallazgos en investigaciones biomédicas y la planificación de experimentos futuros.

Técnicas de Explicabilidad Propuestas

Métodos Post-hoc

- SHAP (SHapley Additive exPlanations): Asigna valores de Shapley a cada característica, explicando la contribución de cada una a las predicciones del modelo.

- LIME (Local Interpretable Model-agnostic Explanations): Proporciona explicaciones locales al ajustar un modelo interpretable a las predicciones de un modelo "caja negra" en la vecindad de una instancia particular.

Modelos Intrínsecamente Interpretables (Ante-hoc)

- Regresiones Lineales: Modelos que proporcionan coeficientes claros que indican la importancia de cada característica.
- Árboles de Decisión: Modelos que presentan una estructura de árbol donde cada nodo representa una decisión basada en una característica, facilitando la interpretación.

Integración de Conocimiento Biológico

- Redes Parcialmente Conectadas Basadas en Anotaciones Biológicas: Incorporan conocimiento biológico conocido en la estructura del modelo, mejorando su interpretabilidad.
- Funciones de Pérdida Basadas en Principios Biológicos: Utilizan funciones de pérdida que reflejan procesos biológicos, haciendo que los modelos sean más interpretables en contextos biomédicos.

Métodos o Técnicas con los que se Compara

Modelos "Caja Negra" Tradicionales

- Redes Neuronales Profundas: Modelos altamente complejos y no transparentes que son difíciles de interpretar sin técnicas de explicación adicionales.
- Máquinas de Soporte Vectorial (SVM): Modelos que separan las clases mediante hiperplanos en un espacio de alta dimensión, cuya interpretación es complicada sin métodos explicativos.

Técnicas Post-hoc Tradicionales

- Análisis de Importancia de Características: Métodos que determinan la importancia de las características de entrada en base a métricas de rendimiento del modelo.
- Visualizaciones de Activación: Técnicas que muestran cómo diferentes partes de una entrada afectan la activación de neuronas específicas en redes neuronales, pero que pueden ser menos precisas o intuitivas que métodos como SHAP y LIME.

La siguiente tabla (véase tabla 43), resume los enfoques más recientes y las mejoras en la explicabilidad de los modelos de aprendizaje automático aplicados a estudios genómicos. En el contexto de la biomedicina, estos avances permiten a los investigadores comprender mejor

los procesos subyacentes y tomar decisiones más informadas. La tabla se organiza en torno a varios aspectos, incluyendo métodos post hoc, modelos intrínsecamente interpretables, y la integración de conocimiento biológico, comparándolos con enfoques tradicionales y destacando su impacto y resultados relevantes en aplicaciones biomédicas.

Tabla 43. Resumen de Enfoques y Mejoras en la Explicabilidad de Modelos de Aprendizaje Automático en Estudios Genómicos.

Aspecto	Descripción
Mejoras en la Explicabilidad	Incluye métodos post hoc y modelos intrínsecamente interpretables, así como la integración de conocimiento biológico para mejorar la comprensión y transparencia de los modelos. Ejemplos de métodos incluyen SHAP, LIME, regresiones lineales, y árboles de decisión.
Marco	Estos métodos y enfoques se agrupan en dos grandes categorías: métodos post hoc (como SHAP y LIME) que proporcionan explicaciones retrospectivas y modelos intrínsecamente interpretables (como regresiones lineales y árboles de decisión) que son diseñados para ser transparentes desde el principio. La integración de conocimiento biológico se utiliza para mejorar la interpretabilidad en contextos biomédicos.
Comparación con Enfoques Anteriores	Los métodos actuales ofrecen una mejor explicación y comprensión en comparación con modelos "caja negra" tradicionales y técnicas post-hoc menos precisas, como el análisis de importancia de características y visualizaciones de activación.
Impacto de las Mejoras	Facilitan la identificación de biomarcadores, permiten pruebas de hipótesis estadísticas clásicas y mejoran la validación y comprensión de los modelos en aplicaciones biomédicas.

Propuestas y Taxonomías	La creación de un espectro de transparencia e interpretabilidad de modelos permite una mejor categorización y comprensión de los niveles de explicabilidad, ayudando en la selección de métodos adecuados para cada aplicación.
Resultados Relevantes	Los modelos interpretables ayudan en la identificación de biomarcadores importantes, la validación de hallazgos en investigaciones biomédicas, y la toma de decisiones clínicas basadas en modelos predictivos precisos y comprensibles.

26. Una revisión sobre la inteligencia artificial explicable en medicina (XAI): Progreso reciente, enfoque de explicabilidad, interacción humana y sistema de puntuación (Sheu & Pardeshi, 2022).

La Inteligencia Artificial Explicable (XAI) está revolucionando la medicina al proporcionar transparencia y comprensión en los modelos predictivos que afectan el diagnóstico y tratamiento de enfermedades. A medida que los modelos de aprendizaje automático se vuelven más sofisticados, surge la necesidad de técnicas que no solo sean precisas, sino también comprensibles para los profesionales de la salud.

Este documento examina los últimos avances en XAI aplicados a la medicina, destacando métodos recientes como enfoques locales y globales, así como técnicas innovadoras como la destilación de conocimiento y el aprendizaje automático interpretable. Se analiza la importancia de la evaluación de estos métodos y su adaptación al contexto médico, incluyendo la interacción humano-computadora (HCI) y el concepto de Human-in-the-Loop (HITL).

Además, se presenta un sistema de puntuación para evaluar la calidad de los sistemas XAI y se discute cómo las técnicas de preprocesamiento y evaluación de datos contribuyen a la efectividad y confiabilidad de estos modelos en la práctica médica. Este análisis busca proporcionar una visión clara de cómo los avances en XAI están mejorando la transparencia y la confianza en la inteligencia artificial aplicada a la salud (Véase tabla 44).

Avances Recientes en Explicabilidad

Métodos Recientes de XAI:

1. **Métodos Locales y Globales:** Se han desarrollado enfoques locales y globales para la explicación de modelos complejos. Estos avances permiten una mayor flexibilidad y especificidad en la explicación de los resultados de los modelos.
2. **Algoritmos Basados en Conocimiento y Destilación:** La incorporación de algoritmos basados en el conocimiento y técnicas de destilación representa un avance significativo al transformar modelos complejos en representaciones más comprensibles sin comprometer la precisión.
3. **Aprendizaje Automático Interpretable:** Se han desarrollado técnicas de aprendizaje automático intrínsecamente interpretables, mejorando así la transparencia y la comprensión de los modelos.

Evaluación y Características de XAI

Evaluación de XAI:

1. **Sistema de Puntuación y Recomendaciones:** El artículo propone un sistema de puntuación para XAI que se basa en la retroalimentación del usuario y la colaboración humano-máquina. Esto es crucial para evaluar la efectividad y utilidad de las explicaciones proporcionadas por los modelos.

Características y Futuro de la Explicabilidad en Salud:

1. **Integración en la Medicina:** Se analiza cómo las características de XAI se relacionan con la explicabilidad en el cuidado de la salud, subrayando la relevancia de estos avances para mejorar la toma de decisiones y la comprensión de las condiciones del paciente.

Comparación con Enfoques Anteriores

1. **Comparación de Métodos:** Se comparan los métodos actuales de XAI con enfoques anteriores, destacando las mejoras en la generación de explicaciones más comprensibles y aplicables a diversos casos de uso. Esto incluye una revisión de cómo las técnicas nuevas superan las limitaciones de métodos anteriores.

Impacto en Propuestas, Taxonomías y Resultados Relevantes

Propuestas y Sistema de Evaluación:

1. **Nuevo Sistema de Recomendación y Puntuación:** Se propone un sistema innovador para la recomendación y evaluación de XAI, lo que representa un avance en la forma de evaluar y validar la explicabilidad de los modelos.

2. **Impacto en la Medicina:** La implementación de soluciones explicables en el ámbito médico subraya la importancia de los avances en XAI para mejorar la toma de decisiones clínicas y la comprensión de las condiciones del paciente.

Futuro de la Explicabilidad en la Atención Sanitaria

1. **Interacción Humano-Computadora (HCI):** La HCI se refiere a la interacción entre el mundo real y la realidad aumentada. En medicina, esto implica registrar interacciones del paciente con sistemas computacionales para identificar síntomas y características. La HCI será clave para la identificación de enfermedades y la mejora de la funcionalidad de los sistemas XAI en la atención sanitaria.
2. **Human-in-the-Loop (HITL):** La integración del concepto HITL es esencial para la aplicación efectiva de XAI en salud. Los expertos deben estar continuamente involucrados en el ajuste y mejora de los modelos interpretables para gestionar los desafíos de datos multimodales y asegurar la confiabilidad del sistema en hospitales.
3. **Evaluación de Explicaciones:** Las explicaciones proporcionadas por los sistemas XAI deben ser efectivas y aceptables para los examinadores médicos. La selección del tipo de explicación más adecuada es importante para la transparencia y la comprensión del sistema de evaluación.
4. **Sistemas Inteligentes Explicables (EIS):** Los avances en XAI están revolucionando la atención sanitaria, desde la robótica asistida en cirugía hasta el descubrimiento de fármacos. Los sistemas inteligentes explicables pueden mejorar el análisis, la predicción y la toma de decisiones, facilitando tratamientos más efectivos y rápidos.

Consideraciones para Aplicar Enfoques de XAI en Datos Médicos y Preprocesamiento

Preprocesamiento de Datos:

1. **Consistencia del Dataset:** Asegurarse de que el dataset esté libre de datos faltantes o erróneos es crucial para evitar errores en las predicciones.
2. **Funciones de Imputación de Datos:** Aplicar técnicas de imputación, como la imputación multivariante por ecuaciones encadenadas (MICE), para corregir datos inconsistentes.
3. **Análisis de Distribuciones de Datos:** Evaluar las distribuciones de datos para entender las relaciones entre observaciones.
4. **Técnicas de Registro de Imágenes:** Alinear correctamente las imágenes médicas para evitar problemas de sesgo y escalado.
5. **Técnicas de Evaluación de Características:** Utilizar métodos como SHAP, GRAD-CAM, y LRP para asignar puntuaciones a las características relevantes.

6. **Prioridad Asignada por Expertos:** Los expertos en el dominio pueden asignar prioridades a características basadas en su experiencia para mejorar la precisión del modelo.
7. **Selección de Características y Umbrales:** Aplicar umbrales y seleccionar características específicas para evitar desequilibrios en el dataset.

Metodología en XAI:

1. **Aspectos de Características para Selección de Métodos:** Elegir métodos de XAI adecuados según el tipo de datos y el algoritmo de aprendizaje.
2. **Genuinidad del Enfoque:** Seleccionar métodos basados en la investigación reciente y realizar experimentos para ajustar los modelos.
3. **Eficiencia de Métodos:** Revisar la literatura reciente para elegir técnicas adecuadas que se adapten al conjunto de datos y al problema.
4. **Análisis de Datos Multi-Modales:** Utilizar modelos híbridos para datos de múltiples formatos y sincronizar modelos para datos de entrada.
5. **Impacto de las Características del Paciente:** Proporcionar detalles sobre el estado del paciente en comparación con datos promedio.
6. **Clasificación del Estado del Paciente:** Asegurarse de que el modelo clasifique correctamente el estado de enfermedad para facilitar el tratamiento.
7. **Curso de Medicación y Resultados:** Monitorear y actualizar los resultados del curso de medicación, especialmente en presencia de comorbilidades.
8. **Porcentaje de Características Clínicas Afectadas:** Proporcionar un resumen detallado del estado clínico del paciente.
9. **Estado de Mejora de Características:** Indicar si las características del paciente muestran mejora o deterioro.
10. **Métricas de Evaluación del Modelo**: Utilizar métricas adecuadas para evaluar el rendimiento del modelo, como precisión y AUROC (Área Bajo la Curva ROC). AUROC, que significa *Area Under the Receiver Operating Characteristic Curve*, es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación, especialmente en problemas de clasificación binaria.

Evaluación en XAI:

1. **Casos Importantes para Clasificación de Salida:** Asegurarse de que el modelo clasifique correctamente los casos binarios o multicategorías.
2. **Mejora Basada en Recursión:** Actualizar el modelo regularmente para mejorar su rendimiento basado en retroalimentación.
3. **Combinación de Salidas Multi-Modelo:** Integrar salidas de diferentes modelos para proporcionar una predicción precisa.

4. **Comparación y Evaluación de Gráficos:** Utilizar gráficos como AUROC y PR para evaluar el rendimiento del modelo.
5. **Selección Manual de Características por Expertos:** Permitir que los expertos seleccionen manualmente características para mejorar la precisión del modelo.
6. **Registro de Retroalimentación de Expertos:** Actualizar los modelos según la retroalimentación de expertos para abordar problemas emergentes.
7. **Actualización de Entrenamiento del Modelo:** Mantener el modelo actualizado para su adaptación y mejora.
8. **Modelo Adecuado para Casos Médicos:** Elegir entre modelos de aprendizaje automático o profundo según la complejidad del caso médico.
9. **Adaptación de Sugerencias de Retroalimentación:** Incorporar sugerencias de retroalimentación para mejorar el desempeño del modelo.

Sistema de Recomendación XAI (XAI-RS)

Objetivo: XAI-RS proporciona recomendaciones personalizadas post-tratamiento para pacientes dados de alta después de recibir tratamiento para diversas enfermedades.

Cómo Funciona:

1. **Evaluación Personalizada:** El sistema evalúa las condiciones de salud recientes del paciente y ofrece recomendaciones personalizadas.
2. **Recomendaciones Generales:** Incluye recomendaciones sobre dieta, medicamentos/tratamientos, ejercicio, chequeos regulares y efectos secundarios. También considera hábitos como fumar o consumir alcohol.

Sistema de Calificación XAI (XAI-SS)

Objetivo: XAI-SS califica la calidad de los sistemas XAI basándose en una serie de factores estandarizados para asegurar su calidad y conformidad.

Cómo Funciona:

1. **Factores de Calificación:** Evalúa 10 factores clave relacionados con XAI, asignando puntos a cada uno basado en criterios específicos.
2. **Clasificación:**
    - Clase I: Calificación $\geq 90\%$
    - Clase II: Calificación $\geq 70\%$ y $< 90\%$
    - Clase III: Calificación $\geq 60\%$ y $< 70\%$

Tabla 44. Enfoque de explicabilidad, interacción humana y sistema de puntuación.

Aspecto	Descripción
Mejoras en la Explicabilidad	Métodos Recientes de XAI: - Métodos Locales y Globales: Se han desarrollado enfoques locales y globales para la explicación de modelos complejos, proporcionando mayor flexibilidad y especificidad. - Algoritmos Basados en Conocimiento y Destilación: La incorporación de algoritmos basados en el conocimiento y técnicas de destilación transforma modelos complejos en representaciones más comprensibles. - Aprendizaje Automático Interpretable: Se han desarrollado técnicas de aprendizaje automático intrínsecamente interpretables, mejorando la transparencia y la comprensión de los modelos.
Marco	La Inteligencia Artificial Explicable (XAI) está revolucionando la medicina al proporcionar transparencia y comprensión en los modelos predictivos que afectan el diagnóstico y tratamiento de enfermedades. A medida que los modelos de aprendizaje automático se vuelven más sofisticados, surge la necesidad de técnicas que no solo sean precisas, sino también comprensibles para los profesionales de la salud.
Comparación con Enfoques Anteriores	Comparación de Métodos: - **Enfoques Tradicionales:** Los métodos recientes de XAI se comparan con técnicas anteriores como árboles de decisión, redes neuronales profundas sin capacidad de interpretación, y técnicas básicas de análisis de características. - **Mejoras identificadas:** Las nuevas técnicas superan las limitaciones de métodos anteriores al ofrecer explicaciones más detalladas y específicas para diferentes casos de uso médico.
Impacto de las Mejoras	La implementación de soluciones explicables en el ámbito médico subraya la importancia de los avances en XAI para mejorar la toma de decisiones clínicas y la comprensión de las condiciones del paciente.
Propuestas y Taxonomías	Propuestas: - **Sistema de Puntuación y Recomendaciones:** Se propone un sistema de puntuación para XAI basado en la

	retroalimentación del usuario y la colaboración humano-máquina. - **Evaluación de XAI:** Se discute cómo las técnicas de preprocesamiento y evaluación de datos contribuyen a la efectividad y confiabilidad de estos modelos en la práctica médica - **Integración en la Medicina:** Se analiza cómo las características de XAI se relacionan con la explicabilidad en el cuidado de la salud, mejorando la toma de decisiones y la comprensión de las condiciones del paciente.
Resultados Relevantes	Este análisis busca proporcionar una visión clara de cómo los avances en XAI están mejorando la transparencia y la confianza en la inteligencia artificial aplicada a la salud.

27. Grafos de Conocimiento como Herramientas para el Aprendizaje Automático Explicable (Tiddi & Schlobach, 2022).

Se propone el uso de grafos de conocimiento (Knowledge Graphs, KGs) como una herramienta eficaz para mejorar la explicabilidad en modelos de aprendizaje automático. Los KGs estructuran y conectan datos diversos, facilitando una comprensión más profunda y transparente de las decisiones de los modelos de machine learning. En comparación con enfoques tradicionales de explicabilidad, como los métodos post-hoc y basados en reglas, los grafos de conocimiento ofrecen ventajas significativas en términos de integración de datos y generación de explicaciones coherentes y comprensibles para los usuarios (véase Tabla 45).

Avances Recientes en la Mejora de la Explicabilidad de los Modelos de Machine Learning 'Caja Negra'.

1. **Integración de Knowledge Graphs (KGs)**:
    - **Descripción**: Los grafos de conocimiento estructuran el conocimiento en relaciones y entidades, lo que permite una navegación y comprensión más avanzadas. Esto facilita que las decisiones de los modelos de machine learning sean más transparentes y comprensibles (Tiddi & Schlobach, 2022).
    - **Precisión**: Correcta. Los KGs permiten una mejor visualización y navegación de datos, haciendo las decisiones de los modelos más comprensibles.
2. **Análisis de Grafos y Estrategias Basadas en Grafos**:
    - **Descripción**: La aplicación de metodologías de análisis de redes y estrategias basadas en grafos ayuda a reducir costos computacionales y mejorar la asignación de recursos al identificar automáticamente la información relevante en los gráficos de conocimiento (Tiddi & Schlobach, 2022).

- **Precisión**: Adecuada. Estas técnicas optimizan el uso de recursos y facilitan la gestión de grandes volúmenes de datos.

3. **Enfoque Híbrido de Inteligencia Artificial**:
   - **Descripción**: La combinación de métodos simbólicos (representación del conocimiento) y no simbólicos (machine learning) ha permitido el desarrollo de sistemas inteligentes híbridos. Estos sistemas explotan los grafos de conocimiento a gran escala para generar explicaciones más coherentes y comprensibles para los humanos (Tiddi & Schlobach, 2022).
   - **Precisión**: Adecuada. La integración de métodos simbólicos y no simbólicos permite un enfoque más robusto para la explicación de decisiones.

Impacto de las Mejoras

1. **Reducción de Costos y Mejor Asignación de Recursos**:
   - **Descripción**: El uso de heurísticas y técnicas avanzadas para la extracción automática de conocimiento de grafos de conocimiento a gran escala ha reducido significativamente los costos computacionales y la necesidad de intervención humana (Tiddi & Schlobach, 2022).
   - **Precisión**: Adecuada. La automatización y el uso de heurísticas optimizan la gestión de datos.
2. **Explicaciones Más Coherentes y Comprensibles**:
   - **Descripción**: La integración de relaciones causales, modales y espacio-temporales en los KGs permite la creación de narrativas complejas que explican eventos de manera más coherente, emulando la comprensión humana. Esto mejora la confianza y la adopción de los sistemas de IA por parte de los usuarios (Tiddi & Schlobach, 2022).
   - **Precisión**: Adecuada. Los KGs facilitan explicaciones más coherentes y comprensibles en comparación con otros métodos.
3. **Enfoque Centrado en el Ser Humano**:
   - **Descripción**: El desarrollo de sistemas de IA centrados en el ser humano, que capturan el significado de experiencias y eventos a través de modelos semánticos, apoya la capacidad humana en lugar de reemplazarla. Esto promueve una IA más colaborativa y ética (Tiddi & Schlobach, 2022).
   - **Precisión**: Adecuada. Este enfoque asegura que la IA complemente y apoye la capacidad humana.

Método o Técnica de Explicabilidad Propuesta y Comparación con Otros Métodos

El paper propone el uso de grafos de conocimiento como una herramienta para mejorar la explicabilidad en machine learning. Comparado con enfoques tradicionales como los

métodos post-hoc y basados en reglas, los KGs ofrecen una ventaja significativa al estructurar el conocimiento de manera flexible y permitir la integración de múltiples fuentes de datos a gran escala. Se destaca la falta de benchmarks y la necesidad de desarrollar criterios de satisfacción y tipos de explicaciones útiles desde una perspectiva humana.

Enfoques Tradicionales:

- **Métodos Post-hoc**: Estos se aplican después de que el modelo ha hecho una predicción para tratar de explicar esa predicción. Ejemplos incluyen LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations). Se centran en proporcionar explicaciones locales específicas para cada predicción.
- **Métodos Basados en Reglas**: Implican la creación de reglas explícitas que pueden ser entendidas y verificadas por humanos, mejorando la transparencia. Sin embargo, estos métodos pueden ser limitados para capturar relaciones complejas y no son fácilmente escalables a grandes conjuntos de datos o modelos complejos.
- **Explicabilidad Mediante Descomposición**: Implica descomponer el modelo en componentes básicos y explicar el papel de cada componente. Este método puede ser útil para modelos simples, pero es ineficaz para modelos complejos debido a la gran cantidad de parámetros e interacciones.

El uso de KGs se presenta como una solución más robusta y escalable, permitiendo la integración de múltiples fuentes de datos y proporcionando explicaciones más coherentes y comprensibles en comparación con los enfoques tradicionales.

Tabla 45. Aspectos de la Explicabilidad mediante Gráficos de Conocimiento en Aprendizaje.

Aspecto	Descripción
Mejoras en la Explicabilidad	Integración de Knowledge Graphs (KGs): Estructuran el conocimiento en relaciones y entidades, facilitando la transparencia y comprensión. Análisis de grafos: Reduce costos computacionales y mejora la asignación de recursos. Enfoque híbrido: Combina métodos simbólicos y no simbólicos para explicaciones más coherentes.
Marco	El marco analítico explora la integración de gráficos de conocimiento en el aprendizaje automático explicable, evaluando cómo estos gráficos estructuran datos diversos para mejorar la comprensión en comparación

	con métodos tradicionales.
Comparación con Enfoques Anteriores	Métodos post-hoc: LIME y SHAP proporcionan explicaciones locales después de la predicción. Métodos basados en reglas: Crean reglas explícitas entendidas por humanos pero limitadas en la captura de relaciones complejas. Explicabilidad mediante descomposición: Efectiva para modelos simples, pero limitada para modelos complejos.
Impacto de las Mejoras	Reducción de costos: Uso de heurísticas y extracción automática de conocimiento disminuye costos y necesidad de intervención humana. Explicaciones coherentes: Mejora la confianza al integrar relaciones causales y modales. Enfoque centrado en el ser humano: Promueve una IA colaborativa y ética.
Propuestas y Taxonomías	Se propone el uso de gráficos de conocimiento para mejorar la explicabilidad en machine learning, destacando su capacidad para integrar múltiples fuentes de datos. Se identifica la necesidad de benchmarks y criterios para evaluar la calidad de las explicaciones desde una perspectiva humana.
Resultados Relevantes	Ventajas de los gráficos de conocimiento: Ofrecen integración flexible de datos y mejores explicaciones en comparación con métodos tradicionales. Desafíos identificados: Falta de benchmarks y métodos eficientes para la extracción y manejo de conocimiento a gran escala.

28. Explicación Local Interpretable Basada en Conceptos con Retroalimentación Humana para Predecir la Mortalidad (Shawi & Al-Mallah, 2022).

El enfoque CLEF (Concept-based Local Explanation Framework) se centra en ofrecer explicaciones locales para la predicción del riesgo de mortalidad, utilizando conceptos definidos por expertos clínicos y explicaciones contrafactuales. Este enfoque permite a los clínicos definir manualmente los conceptos y asociarlos a características específicas, generando explicaciones contrafactuales que muestran qué cambios mínimos en los conceptos serían necesarios para alterar una predicción. La **Tabla 33** proporciona un resumen detallado de cómo CLEF mejora la explicabilidad en comparación con métodos anteriores.

El análisis del impacto de CLEF, tal como se detalla en la **Tabla 33**, destaca su efectividad en la detección de sesgos y la precisión de las explicaciones. CLEF supera a los baselines en

términos de precisión predictiva y en la calidad de las explicaciones, mostrando un impacto positivo en la capacidad de los usuarios para interpretar las predicciones del modelo y detectar sesgos. La **Tabla 46** también ilustra cómo CLEF contribuye a los métodos de explicabilidad basados en conceptos y explicaciones contrafactuales, mejorando la transparencia y comprensión de modelos 'caja negra'.

Detalles del Dataset y Baselines: Para evaluar el enfoque CLEF, se utilizan datos clínicos de pacientes y se comparan los resultados con otros métodos. A continuación se describen los detalles del dataset utilizado y los baselines con los que se compara CLEF.

- **Dataset**: El conjunto de datos proviene del proyecto FIT de Henry Ford e incluye atributos clínicos y vitales de 34,212 pacientes que se sometieron a pruebas de esfuerzo en cinta. Se realizó un seguimiento de 10 años, durante el cual el 11.5% de los pacientes falleció.
- **Conceptos Definidos**: Los conceptos clave son ‘Fitness’, ‘Hypertension’, ‘Obesity/Diabetes’, ‘Dyslipidemia’ y ‘Cardiometabolic’. Cada concepto está asociado a características específicas definidas por los clínicos.
- **Baselines Comparativos**:
    - **Baseline Interactivo (AL)**: Utilizar modelos de regresión logística regularizada y permite la interacción del clínico para explicar los conceptos.
    - **Baseline No Interactivo (LR)**: Basado en regresión logística para proporcionar explicaciones globales del modelo.

Resultados del Estudio: El estudio analiza la efectividad de CLEF en comparación con los baselines, destacando su impacto en la detección de sesgos y la precisión de las explicaciones.

- **Impacto en la Detección de Sesgos**: CLEF muestra una mejora significativa en la capacidad de los usuarios para detectar sesgos en los datos en comparación con los métodos basales.
- **Eficacia de las explicaciones**: Las explicaciones generadas por CLEF son consistentes con el modelo subyacente y permiten una mejor comprensión de las predicciones y los sesgos.
- **Fidelidad de la Explicación**: CLEF proporciona explicaciones alineadas con los cambios en las predicciones del modelo.
- **Confianza en la Explicación**: La confianza en las explicaciones se valida a través de encuestas a expertos, que evalúan la claridad y utilidad de las explicaciones.
- **Detección de sesgos**: CLEF facilita la identificación de sesgos mediante explicaciones contrafactuales, lo que ayuda a interpretar mejor los resultados del modelo.

Relación con la Pregunta de Investigación: Este enfoque aporta avances importantes en la explicabilidad de modelos 'caja negra', integrando la retroalimentación humana y conceptos de alto nivel en las explicaciones.

- **Avances Recientes**: CLEF introduce una metodología interactiva que mejora la explicabilidad en comparación con enfoques que no incorporan interacción humana de manera tan directa.
- **Impacto de las Mejoras**: CLEF demuestra un impacto positivo en la detección de sesgos y la comprensión de las predicciones del modelo, lo que es esencial para la confianza en modelos complejos.
- **Propuestas y Taxonomías**: Este enfoque contribuye a la propuesta de marcos explicativos interactivos y se puede clasificar dentro de los métodos basados en conceptos y explicaciones contrafactuales.

Construcción de Explicaciones Locales (CLEF): CLEF utiliza dos modelos para proporcionar explicaciones locales y contrafactuales para las predicciones.

- **Árbol de Decisión**: Genera contrafactuales mostrando qué cambios en los atributos llevarían a una predicción diferente.
- **Regresión Logística**: Ajusta los conceptos mediante pesos y proporciona contrafactuales al modificar los valores de los conceptos para cambiar la predicción.

Resultados y Evaluación: Se evaluó el desempeño de CLEF utilizando el dataset del proyecto FIT y se comparó con los baselines establecidos.

- **Conjunto de Datos**: Incluye atributos clínicos y vitales como edad, presión arterial y enfermedades.
- **Definición de Conceptos**: Conceptos definidos como 'Fitness', 'Hipertensión', 'Obesidad/Diabetes', 'Dislipidemia' y 'Cardiometabólico' con características específicas asociadas.
- **Comparación de Baselines**:
    - **Baseline Interactivo (AL)**: Utilizar clasificadores de conceptos para explicar predicciones.
    - **Baseline No Interactivo (LR)**: Basado en regresión logística regularizada para explicaciones globales.
- **Resultados**: CLEF supera a ambos baselines en precisión predictiva y en la calidad de las explicaciones.
- **Fidelidad de la Explicación**: Las explicaciones generadas por CLEF son coherentes con los cambios en las predicciones.

- **Confianza en la explicación**: La confianza en las explicaciones se valida a través de encuestas a expertos.
- **Detección de sesgos**: CLEF ayuda a identificar sesgos en los datos mediante explicaciones contrafactuales, mejorando la interpretabilidad del modelo.

Tabla 46. Avances y Comparaciones del Enfoque CLEF para Explicabilidad Local en Modelos Predictivos.

Aspecto	Descripción
Mejoras en la Explicabilidad	CLEF mejora la explicabilidad al ofrecer explicaciones locales basadas en conceptos definidos por expertos clínicos y explicaciones contrafactuales que muestran los cambios mínimos necesarios para alterar una predicción.
Marco	CLEF (Concept-based Local Explanation Framework) es un enfoque que proporciona explicaciones locales utilizando conceptos de alto nivel definidos por humanos y generando contrafactuales para facilitar la interpretación.
Comparación con Enfoques Anteriores	CLEF se compara con dos baselines: un enfoque interactivo (AL) que utiliza regresión logística regularizada y permite interacción con el clínico, y un enfoque no interactivo (LR) basado en regresión logística global.
Impacto de las Mejoras	CLEF tiene un impacto positivo en la capacidad de los usuarios para detectar sesgos y entender las predicciones del modelo. Las explicaciones generadas son coherentes y ayudan a interpretar mejor los resultados del modelo.
Propuestas y Taxonomías	CLEF contribuye a los métodos de explicabilidad basados en conceptos y explicaciones contrafactuales. Ofrece un enfoque interactivo para mejorar la transparencia y comprensión de modelos 'caja negra'.
Resultados Relevantes	CLEF supera a los baselines en precisión predictiva y en la calidad de las explicaciones. La fidelidad de las explicaciones y la detección de sesgos son significativamente mejoradas en comparación con los enfoques tradicionales.

29. Sobre la Explicabilidad de los Modelos Profundos de Procesamiento de Lenguaje Natural (El Zini & Awad, 2023).

Tradicionalmente, los enfoques post-hoc como LIME y SHAP han dominado la escena, proporcionando interpretaciones locales y explicaciones aproximadas después de que el modelo ha sido entrenado. Sin embargo, estos métodos, aunque útiles, presentan limitaciones en términos de coherencia y robustez de las explicaciones.

Recientemente, se han desarrollado nuevas metodologías que integran la explicabilidad directamente en el proceso de aprendizaje del modelo, ofreciendo interpretaciones más coherentes y robustas. Herramientas innovadoras como exBERT y benchmarks como ERASER están redefiniendo los estándares para evaluar y mejorar la explicabilidad en el procesamiento del lenguaje natural (NLP). Además, enfoques basados en la teoría de la información y la interpretación de embeddings de palabras están proporcionando una visión más profunda y estructurada de las representaciones aprendidas por los modelos de machine learning.

El documento presenta un análisis detallado de estos avances recientes, comparándolos con enfoques anteriores y evaluando su impacto en términos de propuestas, taxonomías y resultados relevantes. A través de estudios de caso y ejemplos prácticos, se exploran las nuevas técnicas que están llevando la explicabilidad de los modelos de machine learning a un nuevo nivel, facilitando una mayor confianza y transparencia en aplicaciones críticas.

Clasifica los métodos de explicabilidad en tres niveles:

Nivel de Entrada: Explicaciones sobre incrustaciones de palabras y su impacto en los modelos.

Nivel de Procesamiento: Explicaciones sobre cómo los modelos de NLP procesan la información textual internamente.

Nivel de Salida: Explicaciones sobre las decisiones y predicciones generadas por los modelos.

Esta clasificación ayuda a abordar distintos aspectos de la interpretabilidad y proporciona un marco claro para aplicar y evaluar métodos de explicación.

Avances Recientes en la Mejora de la Explicabilidad:

1. **Descomposición de Redes Neuronales y Visualización**:

- **exBERT** (Hoover et al., 2019): Esta herramienta permite descomponer y visualizar las capas de redes neuronales transformadoras. Proporciona una interfaz interactiva que permite a los usuarios explorar las representaciones aprendidas y entender mejor el comportamiento del modelo.
- **Visualización y Manipulación Interactiva de la Atención** (Lee et al., 2017): En el contexto de la traducción automática basada en redes neuronales, este enfoque permite visualizar cómo el modelo distribuye su atención durante la traducción. Ayuda a los investigadores y usuarios a comprender las relaciones de dependencia entre palabras en diferentes idiomas.

2. **Evaluación y Benchmarking de Modelos**:
   - **ERASER** (DeYoung et al., 2020): Es un conjunto de datos y un marco para evaluar modelos racionalizados en procesamiento de lenguaje natural. Proporciona métricas específicas para medir la calidad y la utilidad de las explicaciones generadas por los modelos, lo que facilita la comparación de diferentes enfoques de interpretabilidad.
   - **Medición de la Interpretabilidad de Modelos** (Doshi-Velez & Kim, 2017): Propone un marco para evaluar la interpretabilidad de los modelos de machine learning, considerando aspectos como la comprensibilidad para los humanos y la utilidad práctica de las explicaciones.
3. **Interpretabilidad de Embeddings de Palabras**:
   - **Embeddings en el Dominio Médico** (Jha et al., 2018): Este estudio analiza cómo los embeddings de palabras capturan información sintáctica y semántica en textos médicos, lo que permite una mejor interpretación de las representaciones aprendidas y su aplicación en tareas médicas.
   - **Aspectos Sintácticos y Semánticos en Embeddings** (Kádár et al., 2017): Analiza cómo los embeddings capturan diferentes aspectos del lenguaje, proporcionando una visión más profunda de cómo los modelos de NLP comprenden y representan el texto.
4. **Metodologías de Interpretación Basadas en Teoría de la Información**:
   - **L-Shapley y C-Shapley** (Chen et al., 2018): Estos métodos utilizan conceptos de la teoría de la información para interpretar datos estructurados. Ofrecen una manera eficiente de entender la importancia de las características en el proceso de toma de decisiones del modelo.

Comparación con Enfoques Anteriores:

- **Métodos Post-Hoc**:

- **LIME** (Ribeiro et al., 2016): Este método proporciona explicaciones locales para las predicciones de cualquier modelo. Sin embargo, sus explicaciones son independientes del proceso de entrenamiento del modelo.
- **SHAP** (Lundberg & Lee, 2017): Utiliza valores de Shapley de la teoría de juegos para asignar la importancia a cada característica en una predicción dada. Aunque es más consistente y teóricamente sólido que LIME, sigue siendo un enfoque post-hoc.

- **Integración de la Explicabilidad en el Modelo**:
  - **Redes Neuronales Modulares** (Hu et al., 2017): Proponen una arquitectura modular que incorpora la explicabilidad en el propio diseño del modelo, facilitando interpretaciones más coherentes y robustas.
  - **Racionalización Invariante** (Chang et al., 2020): Este enfoque asegura que las explicaciones sean consistentes y robustas frente a diferentes variaciones de los datos, integrando la interpretabilidad en el proceso de aprendizaje.

Impacto de las Mejoras:

1. **Propuestas**:
   - Herramientas como **exBERT** y benchmarks como **ERASER** están estableciendo nuevos estándares en la evaluación de la explicabilidad en NLP. Estas propuestas no solo mejoran la comprensión de los modelos, sino que también facilitan la comparación y la mejora continua de las técnicas de explicabilidad.
2. **Taxonomías**:
   - Las nuevas taxonomías ayudan a clasificar las diferentes dimensiones y características de la interpretabilidad, proporcionando un marco más estructurado para evaluar y comparar modelos. Por ejemplo, la clasificación de Peters et al. (2018) sobre la calidad de los embeddings permite una evaluación más granular de cómo los modelos de NLP representan el lenguaje.
3. **Resultados Relevantes**:
   - La integración de la explicabilidad directamente en los modelos ha resultado en una mayor confianza y transparencia en aplicaciones críticas. En el dominio médico, por ejemplo, los avances en interpretabilidad han facilitado la adopción de modelos de IA en el diagnóstico y tratamiento, mejorando la toma de decisiones clínicas y la confianza de los profesionales de la salud en estas tecnologías.

Ejemplos:

El documento proporciona estudios de caso y ejemplos detallados que ilustran cómo se han aplicado estas mejoras en diferentes contextos. Por ejemplo:

- **Interpretación de Modelos de NLP**: Se discuten varias herramientas y técnicas para visualizar y entender modelos de lenguaje, destacando cómo estas contribuyen a una mejor comprensión de los procesos internos de los modelos y a una toma de decisiones más informada.
- **Evaluación de la Calidad de las explicaciones**: Se mencionan diferentes métricas y enfoques para evaluar la calidad de las explicaciones generadas por los modelos, subrayando la importancia de tener criterios claros y estandarizados para medir la interpretabilidad.

Como se resume en la Tabla 47, los avances recientes en la explicabilidad de los modelos profundos de procesamiento de lenguaje natural incluyen nuevas metodologías que integran la explicabilidad directamente en el proceso de aprendizaje del modelo, como exBERT y ERASER. La tabla también clasifica los métodos de explicabilidad en tres niveles: entrada, procesamiento y salida, y compara estos enfoques con los métodos post-hoc tradicionales como LIME y SHAP. Además, la tabla destaca el impacto de estas mejoras en términos de propuestas, taxonomías y resultados relevantes, mostrando cómo estas innovaciones están estableciendo nuevos estándares y facilitando una mayor confianza y transparencia en aplicaciones críticas.

Tabla 47. Resumen de Avances Recientes en la Explicabilidad de Modelos Profundos de Procesamiento de Lenguaje Natural.

Aspecto	Descripción
Mejoras en la Explicabilidad	- **Métodos Integrados**: Nuevas metodologías como exBERT y enfoques de interpretación de embeddings. - **Evaluación y Benchmarking**: ERASER para medir la calidad de las explicaciones. - **Teoría de la Información**: L-Shapley y C-Shapley para interpretar datos estructurados.
Marco	- Clasificación en tres niveles: 1. **Entrada**: Explicaciones sobre embeddings de palabras. 2. **Procesamiento**: Cómo los modelos procesan la información. 3. **Salida**: Explicaciones sobre decisiones del modelo.
Comparación con Enfoques Anteriores	- **Métodos Post-Hoc**: LIME y SHAP para explicaciones locales, aunque presentan limitaciones en coherencia y robustez.

	- **Integración en el Modelo**: Redes neuronales modulares y racionalización invariante para explicaciones más coherentes y robustas.
Impacto de las Mejoras	- **Propuestas**: Nuevas herramientas y benchmarks establecen nuevos estándares en evaluación y mejora de la explicabilidad. - **Taxonomías**: Nuevas taxonomías estructuran y clasifican la interpretabilidad de modelos. - **Resultados Relevantes**: Mayor confianza y transparencia en aplicaciones críticas como el diagnóstico médico.
Propuestas y Taxonomías	- **Propuestas**: exBERT, ERASER, herramientas de visualización interactiva. - **Taxonomías**: Clasificación de Peters et al. (2018) sobre embeddings y calidad de explicaciones.
Resultados Relevantes	- **Aplicaciones**: Ejemplos en interpretación de modelos de NLP y evaluación de la calidad de las explicaciones. - **Estudios de Caso**: Aplicación de nuevas técnicas en contextos prácticos y evaluación estandarizada.

30. Revisión Integral y Aplicación de un Modelo de Aprendizaje Profundo Interpretable para la Predicción de Reacciones Adversas a los Medicamentos (ADR) (Dubey & Pandit, 2022).

La seguridad de los medicamentos es un pilar importante en la atención médica moderna. Las reacciones adversas a los medicamentos (ADRs) representan efectos no deseados y perjudiciales que pueden variar desde molestias leves hasta riesgos graves para la salud, incluyendo hospitalización y, en algunos casos, la muerte. A pesar de los rigurosos procesos de pruebas clínicas, los medicamentos aprobados aún pueden presentar riesgos significativos, lo que subraya la necesidad de métodos efectivos para predecir y detectar ADRs tempranamente.

El estudio tiene como objetivo diseñar y desarrollar un modelo de aprendizaje profundo interpretable para la predicción de ADRs. La falta de interpretabilidad en los modelos de aprendizaje profundo tradicionales ha sido abordada mediante la implementación de técnicas de explicación local como LIME (Local Interpretable Model-Agnostic Explanations) y SP-LIME (Sub-modular Pick Local Interpretable Model-Agnostic Explanations), que permiten una comprensión más clara y transparente de las predicciones del modelo.

Además, el estudio evalúa la eficacia del modelo propuesto comparándolo con enfoques anteriores, destacando las mejoras en términos de precisión, confiabilidad y utilidad clínica.

Los resultados obtenidos demuestran que el uso de modelos interpretables no solo mejora la capacidad de predicción, sino que también facilita la aceptación y el uso práctico en entornos médicos, contribuyendo así a una mayor seguridad del paciente y reducción de los riesgos asociados con el uso de medicamentos.

Avances Recientes en la Explicabilidad

1. Uso de Técnicas Interpretable Model-Agnostic

El paper destaca la aplicación de LIME (Local Interpretable Model-Agnostic Explanations) y SP-LIME (Sub-modular Pick Local Interpretable Model-Agnostic Explanations) como avances significativos en la mejora de la explicabilidad de los modelos de aprendizaje profundo. Estas técnicas permiten generar explicaciones locales para predicciones individuales y seleccionar subconjuntos de datos representativos para entender el comportamiento global del modelo.

2. Implementación de Modelos de Sustitución

El paper propone el uso de modelos de sustitución interpretable para abordar la falta de interpretabilidad en los modelos de redes neuronales profundas (DNN). Estos modelos de sustitución ayudan a proporcionar una explicación comprensible del comportamiento de los modelos más complejos, facilitando su aceptación y confianza en el dominio médico.

Comparación con Enfoques Anteriores

Enfoques Tradicionales

Anteriormente, la detección y predicción de ADRs se realizaba utilizando técnicas estadísticas y de minería de datos, como el Análisis de Desproporcionalidad (DPA) y las técnicas de Procesamiento de Lenguaje Natural (NLP). Estos enfoques, aunque útiles, carecían de la capacidad de proporcionar explicaciones detalladas y locales sobre las predicciones.

Modelos de Aprendizaje Automático Supervisado

Las técnicas supervisadas como SVM, Árboles de Decisión y KNN ofrecían cierto nivel de interpretabilidad, pero no eran adecuadas para capturar relaciones complejas en los datos de ADRs. Además, la interpretabilidad era limitada a las características utilizadas en los modelos.

Impacto de las Mejoras

Propuestas y Taxonomías

Las mejoras en la explicabilidad, como el uso de LIME y SP-LIME, han permitido el desarrollo de modelos más comprensibles y transparentes. Estas técnicas han facilitado la creación de taxonomías más detalladas y precisas para la clasificación y predicción de ADRs, mejorando la capacidad de los investigadores y profesionales de la salud para entender y confiar en los modelos.

Resultados Relevantes

El impacto de estas mejoras se refleja en la capacidad de los modelos para proporcionar explicaciones detalladas y comprensibles sobre las predicciones, lo que ha llevado a una mayor aceptación y uso de modelos de aprendizaje profundo en el dominio médico. La integración de datos validados de asociaciones fármaco-ADR ha mejorado la precisión y fiabilidad de las predicciones, contribuyendo a una mejor seguridad del paciente y reducción de riesgos asociados con la medicación.

Para una visión más detallada de las mejoras y comparaciones entre diferentes enfoques, se puede consultar la Tabla 48.

Tabla 48.Avances Recientes en la Explicabilidad de Modelos de Aprendizaje Profundo para la Predicción de Reacciones Adversas a Medicamentos: Un Enfoque Comparativo.

Aspecto	Descripción
Mejoras en la Explicabilidad	Aplicación de técnicas interpretable model-agnostic como LIME y SP-LIME para generar explicaciones locales y entender el comportamiento global del modelo. Implementación de modelos de sustitución interpretables para explicar modelos de redes neuronales profundas.
Marco	Técnicas Interpretable Model-Agnostic. Modelos de Sustitución Interpretable.
Comparación con Enfoques Anteriores	**Enfoques Tradicionales**: ● **Análisis de Desproporcionalidad (DPA)**: Utilizado para identificar señales de ADRs en bases de datos de farmacovigilancia. ● **Técnicas de Procesamiento de Lenguaje Natural (NLP)**: Empleadas para extraer información relevante sobre ADRs de textos médicos y reportes.

	Modelos de Aprendizaje Automático Supervisado: ● **Máquinas de Soporte Vectorial (SVM)**: Proporcionan cierta interpretabilidad, pero no son ideales para captar relaciones complejas en los datos. ● **Árboles de Decisión**: Ofrecen una forma de interpretabilidad mediante sus reglas de decisión, pero pueden no capturar adecuadamente las interacciones complejas. ● **K-Nearest Neighbors (KNN)**: Ofrece interpretabilidad basada en la proximidad de los datos, pero tiene limitaciones para modelar relaciones complejas en datos de ADRs.
Impacto de las Mejoras	Mejora en la precisión y fiabilidad de las predicciones. Mayor aceptación y uso de modelos de aprendizaje profundo en el ámbito médico, contribuyendo a una mejor seguridad del paciente y reducción de riesgos asociados con la medicación.
Propuestas y Taxonomías	Desarrollo de taxonomías más detalladas y precisas para la clasificación y predicción de ADRs, facilitando la comprensión y confianza en los modelos de aprendizaje profundo.
Resultados Relevantes	La integración de datos validados y técnicas avanzadas de explicación ha llevado a una mayor precisión en las predicciones, aceptación en entornos médicos y una reducción en los riesgos asociados con medicamentos.

Anexo II

Artículos Recuperados sobre Modelos de Explicabilidad en Inteligencia Artificial

Introducción

En este capítulo se presentan todos los artículos recuperados de bases de datos académicas como ACM Digital Library, Google Scholar e IEEE Xplore. La selección de estos artículos se basó en el uso de palabras clave siguiendo el enfoque PICO (ver sección 3.2). Se incluyen tanto aquellos que cumplen completamente con la pregunta de investigación planteada como aquellos que no la cumplen.

Cada entrada de la tabla proporciona información sobre el título del artículo, la cita correspondiente y comentarios que destacan los aspectos relevantes de cada estudio.

Papers recuperados de ACM Digital Library, Google Scholar, IEEE *Xplore*.

Nº	Título del paper	Cita	Comentarios
1	A regional explanation for Laxfordian tectonic evolution and its implications for the Lewisian terrane model.	Graham Park; A regional explanation for Laxfordian tectonic evolution and its implications for the Lewisian terrane model.(2022) Scottish Journal of Geology;; 58 (1): sjg2021–020. doi: https://doi.org/10.1144/sjg2021-020	El abstract describe la historia estructural y metamórfica tardía del Complejo Lewisiano durante el Paleoproterozoico, específicamente durante el período Laxfordiano.
2	Alternative interpretable machine learning models applied to corporate probability of default: A literature review and high points of a benchmarking analysis	Jacobs, M. (2024). Alternative interpretable machine learning models applied to corporate probability of default: A literature review and high points of a benchmarking analysis. *SSRN*. https://doi.org/10.2139/ssrn.4583014	El abstract no cumple completamente con la pregunta de investigación planteada porque: No proporciona información clara sobre nuevas metodologías recientes desarrolladas para la explicabilidad de modelos 'caja negra'. Sí realiza una comparación con enfoques anteriores, lo que es positivo y relevante para la pregunta de investigación. No menciona propuestas nuevas, taxonomías, frameworks específicos, ni presenta estudios de caso que muestren el impacto práctico de las mejoras en la explicabilidad.
3	Clustering-based cancer diagnosis model for whole slide image.	Shakeel Sheikh, T., Shim, J., & Cho, M. (2023). Clustering-based cancer diagnosis model for whole slide image. In Proceedings of the 2023 8th International Conference on Biomedical Imaging, Signal Processing (ICBSP '23) (pp. 1–8). https://doi.org/10.1145/3634875.3634876	Este paper introduce una técnica innovadora llamada Clustering-Based Cancer Diagnosis (CBCD) para la clasificación de Whole Slide Images (WSIs) en subtipos de cáncer, mejorando tanto el rendimiento como la interpretabilidad en comparación con los métodos del estado del arte (SOTA). El CBCD utiliza técnicas de agrupamiento como k-means, modelo de mezcla gaussiana y agrupamiento aglomerativo, y se evalúa mediante métricas como adjusted rand y calinski harabasz scores. Se demuestra que el CBCD logra un mejor rendimiento y mayor interpretabilidad en la clasificación de subtipos de cáncer en WSIs.

4	Computational philosophy	Etienne, H. (2022). Computational philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)* (p. 899). https://doi.org/10.1145/3514094.3539562	La "filosofía computacional" no es un método de inteligencia artificial en sí misma. Es un enfoque interdisciplinario que combina principios filosóficos con técnicas computacionales para abordar problemas complejos en campos como las interacciones sociales en línea.
5 ACM	Contrastive counterfactual fairness in algorithmic decision-making	Mutlu, E. Ç., Yousefi, N., & Garibay, O. O. (2022). Contrastive counterfactual fairness in algorithmic decision-making. En Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22) (pp. 499–507). https://doi.org/10.1145/3514094.3534143	El paper aborda la pregunta de investigación al describir avances recientes en la mejora de la explicabilidad y equidad de los modelos de machine learning. Se introduce un enfoque novedoso y generalizable para abordar la causalidad y equidad, comparándolo con enfoques anteriores y mostrando un impacto claro y positivo en términos de propuestas innovadoras, nuevas taxonomías de equidad y resultados empíricos sólidos.
6 ACM	Efficient IoT Traffic Inference: From Multi-view Classification to Progressive Monitoring	Pashamokhtari, A., Batista, G., & Gharakheili, H. H. (Year). Efficient IoT Traffic Inference: From Multi-view Classification to Progressive Monitoring. ACM Transactions on Internet of Things, 5(1), 1–30. https://doi.org/10.1145/3625306	El abstract resalta el potencial impacto de las mejoras propuestas en la inferencia del comportamiento de la red IoT, incluida la reducción del costo de procesamiento y el mantenimiento de altos niveles de precisión. Se destaca la liberación pública de datos y código para facilitar la replicabilidad y la adopción de las propuestas.
7 ACM	Ethical Implications of Transparency and Explainability of Artificial Intelligence for Managing Value-Added Tax (VAT) in Corporations.	Yordanova, Z. (2024). Ethical Implications of Transparency and Explainability of Artificial Intelligence for Managing Value-Added Tax (VAT) in Corporations. En T. Guarda, F. Portela y J.M. Diaz-Nafria (Eds.), Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2023. (Comunicaciones en Ciencia de la Computación e Información, Vol. 1936). Springer, Cham. https://doi.org/10.1007/978-3-031-48855-9_26	El resumen concluye que la investigación contribuirá al conocimiento en el campo de la tecnología fiscal y la AI, proporcionando una visión integral de las soluciones de AI para la gestión del IVA y las perspectivas de las partes interesadas.
8 ACM	Explainability in Mechanism Design: Recent Advances and the Road Ahead.	Suryanarayana, S.A., Sarne, D., Kraus, S. (2022). Explainability in Mechanism Design: Recent Advances and the Road Ahead. In: Baumeister, D., Rothe, J. (eds) Multi-Agent Systems. EUMAS 2022. Lecture Notes in Computer Science(), vol 13442. Springer, Cham.	Se revisa la explicabilidad en el diseño de mecanismos, un campo donde los agentes toman decisiones económicas y no siempre hay una opción que maximice todas las utilidades individuales. Aunque gran parte del enfoque ha sido en explicar algoritmos de caja negra, la explicación de decisiones en mecanismos sociales está ganando relevancia. Se discuten las propiedades y objetivos específicos de la explicabilidad en este contexto y se revisan los principales desafíos y posibles soluciones en el

		https://doi.org/10.1007/978-3-031-20614-6_21	diseño de mecanismos explicables.
9 ACM	Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence.	Baker, S., & Xiang, W. (2023). Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence. *Journal Name*. Retrieved from https://www.researchgate.net/profile/Stephanie-Baker-12/publication/376412417_Explainable_AI_is_Responsible_AI_How_Explainability_Creates_Trustworthy_and_Socially_Responsible_Artificial_Intelligence/links/6577d1cdfc4b416622b8b444/Explainable-AI-is-Responsible-AI-How-Explainability-Creates-Trustworthy-and-Socially-Responsible-Artificial-Intelligence.pdf	No se identificaron nuevas metodologías recientes específicas para mejorar la explicabilidad de los modelos 'caja negra'. No proporciona una comparación explícita con métodos anteriores en términos de explicabilidad. Parcialmente menciona propuestas nuevas sobre la integración de XAI con RAI, pero no detalla frameworks o taxonomías específicas. No presenta resultados prácticos, estudios de caso ni otros resultados específicos que demuestren el impacto práctico de las mejoras en la explicabilidad.
10 ACM	Exploiting negative preference in content-based music recommendation with contrastive learning.	Park, M., & Lee, K. (2022). Exploiting negative preference in content-based music recommendation with contrastive learning. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22)* (pp. 229–236). https://doi.org/10.1145/3523227.3546768	Se centra en analizar el papel de las preferencias negativas en los gustos musicales de los usuarios y en evaluar diferentes estrategias de entrenamiento para un sistema de recomendación de música.
11 ACM	Federated Explainability for Network Anomaly Characterization	Sáez-de-Cámara, X., Flores, J. L., Arellano, C., Urbieta, A., & Zurutuza, U. (2023). Federated Explainability for Network Anomaly Characterization. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '23)* (pp. 346–365). https://doi.org/10.1145/3607199.3607234	El documento presenta un enfoque para mejorar la explicabilidad en sistemas de detección de intrusiones basados en ML, especialmente en entornos distribuidos como el Internet de las cosas (IoT). Destaca la importancia de proporcionar información contextual para que los analistas de seguridad comprendan por qué se clasificó una muestra como anómala y cómo correlacionar diferentes tipos de anomalías. La metodología propuesta adapta algoritmos de explicabilidad, agrupamiento y validación de grupos para extraer patrones en muestras anómalas y identificar amenazas en toda la red. Los resultados demuestran la utilidad de este enfoque en conjuntos de datos de detección de intrusiones del mundo real.
12 ACM	Hybrid Explainable Intrusion Detection System: Global vs.	Tanuwidjaja, H. C., Takahashi, T., Lin, T.-N., Lee, B., & Ban, T.	Este estudio propone un sistema de detección de intrusiones explicables (X-IDS) para mejorar la comprensión de los operadores de seguridad en sistemas

	Local Approach	(2023). Hybrid explainable intrusion detection system: Global vs. local approach. In *Proceedings of the 2023 Workshop on Recent Advances in Resilient and Trustworthy ML Systems in Autonomous Networks (ARTMAN '23)* (pp. 37–42). https://doi.org/10.1145/3605772.3624004	basados en aprendizaje automático. Se aplican técnicas de Inteligencia Artificial Explicable (XAI), como interpretaciones de modelos locales y explicaciones aditivas de Shapley. Se desarrolla un marco de explicación que incluye gráficos de importancia de variables, gráficos de valores individuales y gráficos de dependencia parcial. Se sugiere investigar técnicas adicionales de explicabilidad y evaluar su impacto en la eficacia del sistema de detección de intrusiones.
13 ACM	Identifying Bias in Data Using Two-Distribution Hypothesis Tests	Yik, W., Serafini, L., Lindsey, T., & Montañez, G. D. (2022). Identifying Bias in Data Using Two-Distribution Hypothesis Tests. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22) (pp. 831–844). https://doi.org/10.1145/3514094.3534169	Presenta un avance reciente en la mejora de la explicabilidad de los modelos de machine learning considerados 'caja negra' al introducir un enfoque novedoso para identificar y mitigar sesgos en conjuntos de datos de entrenamiento. Además, destaca el impacto de estas mejoras al proporcionar explicaciones plausibles para los sesgos encontrados, lo que permite una comprensión más profunda de los procesos subyacentes que generaron los datos.
14 ACM	Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank.	Sipos, L., Schäfer, U., Glinka, K., & Müller-Birn, C. (2023). Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank. In *Mensch und Computer 2023 (MuC '23)*, September 03–06, 2023, Rapperswil, Switzerland (pp. 1–6). ACM, New York, NY, USA. https://doi.org/10.1145/3603555.3608851	Este estudio se centra en el método de explicabilidad llamado XAI Question Bank (XAIQB), el cual consiste en un conjunto de preguntas diseñadas para que los usuarios finales, como expertos en la materia y usuarios comunes, las planteen al interactuar con sistemas de inteligencia artificial (IA). El objetivo principal del XAIQB es ayudar a los desarrolladores y diseñadores a comprender y abordar las necesidades de explicación de los usuarios al utilizar sistemas de IA. El paper utiliza el XAIQB para analizar 12 exploraciones de software realizadas por historiadores del arte, con el propósito de evaluar su eficacia para identificar las necesidades de explicación en un contexto específico. Como resultado de este análisis, se amplía el XAIQB con 11 nuevas preguntas y se mejoran las descripciones de todas las preguntas existentes para facilitar su aplicación y comprensión.
15 ACM	Interaction Techniques with a Navigation Robot for the Visually Impaired	Asakawa, C. (2023). Interaction Techniques with a Navigation Robot for the Visually Impaired. In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23) (p. 1). https://doi.org/10.1145/3568162.3576952	El abstract no se centra explícitamente en técnicas innovadoras para mejorar la explicabilidad de modelos 'caja negra' en inteligencia artificial.Describe el desarrollo de un robot de navegación para personas con discapacidad visual utilizando tecnología avanzada, pero no se mencionan métodos específicos para hacer que los modelos de inteligencia artificial sean más explicables.
16	Machine Learning	Balla, Y., Tirunagari, S., &	Se menciona que el surgimiento de herramientas de

ACM	in Pediatrics: Evaluating Challenges, Opportunities, and Explainability.	Windridge, D. (2023, May 14). Machine Learning in Pediatrics: Evaluating Challenges, Opportunities, and Explainability. *PMID: 37179470*. Advance online publication. https://doi.org/S097475591600533	Inteligencia Artificial (IA) como ChatGPT y Bard está provocando cambios significativos en diversos campos, incluida la medicina. En la medicina pediátrica, la IA también se está utilizando cada vez más en múltiples subespecialidades. Sin embargo, la aplicación práctica de la IA aún enfrenta varios desafíos clave. Por lo tanto, hay una necesidad de obtener una visión concisa de los roles de la IA en los múltiples dominios de la medicina pediátrica, que es el objetivo del estudio actual.
17 ACM	MetaFraming: A Methodology for Democratizing Heritage Interpretation Through Wiki Surveys	Wehmeier, C., & Artopoulos, G. (2023). MetaFraming: A Methodology for Democratizing Heritage Interpretation Through Wiki Surveys. In Proceedings of the 20th International Conference on Culture and Computer Science: Code and Materiality (KUI '23) (Article No. 4, pp. 1–9). https://doi.org/10.1145/3623462.3623465	El desarrollo de MetaFraming implica el uso de herramientas computacionales y encuestas wiki para abordar los desafíos en la interpretación del patrimonio cultural. Comienza con la creación de datos semánticamente estructurados a partir de notas de investigación mediante IA. Estos datos se utilizan para iniciar una encuesta wiki colaborativa, donde los participantes pueden clasificar propuestas, comentar y contribuir con nuevas ideas. Además, se utiliza un proceso automatizado para reconstruir el contexto de las acciones de los participantes. Este enfoque fomenta la transparencia y la participación en la interpretación del patrimonio cultural al permitir que los participantes expresen sus perspectivas de manera abierta y colaborativa.
18 ACM	Modern Theoretical Tools for Understanding and Designing Next-generation Information Retrieval System	Xu, D., & Ruan, C. (Year). Modern Theoretical Tools for Understanding and Designing Next-generation Information Retrieval System. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), 1635–1637. https://doi.org/10.1145/3488560.3501394	El abstract presenta una visión crítica sobre el equilibrio entre el progreso ingenieril y teórico en el aprendizaje automático, particularmente en el contexto de la recuperación de información (IR). Destaca que, si bien la última ola de inteligencia artificial ha traído técnicas poderosas, el equilibrio se está inclinando hacia la aplicación, lo que ha generado dependencia de mecanismos de "caja negra". Se propone abordar esta brecha mediante la adopción de herramientas teóricas avanzadas y se ofrece un tutorial sistemático sobre cómo adaptarlas para resolver problemas modernos de IR (recuperación de información)
19 ACM	Multi-view contrastive self-supervised learning of accounting data representations for downstream audit tasks	Schreyer, M., Sattarov, T., & Borth, D. (2021). Multi-view contrastive self-supervised learning of accounting data representations for downstream audit tasks. In *Proceedings of the Second ACM International Conference on AI in Finance (ICAIF '21)* (Article No. 8, pp. 1–8). https://doi.org/10.1145/3490354.3494373	El método de explicabilidad propuesto para auditoría es un marco de aprendizaje auto-supervisado contrastivo. Este marco está diseñado para aprender representaciones de datos contables que sean invariantes a las tareas de auditoría. Utiliza políticas deliberadas de aumento de datos basadas en los atributos de los datos de asientos contables. Este enfoque busca mejorar la eficiencia de la auditoría al proporcionar representaciones ricas e interpretables que sean adecuadas para múltiples tareas de auditoría.
20 ACM	MurTree: optimal decision trees via Dynamic	Demirović, E., Lukina, A., Hebrard, E., Chan, J., Bailey, J., Leckie, C., Ramamohanarao, K., & Stuckey, P.	A pesar de ser un abstract que presenta avances significativos en la optimización de árboles de decisión, no cumple completamente con los criterios de inclusión

	programming and search	J. (2022). MurTree: Optimal decision trees via Dynamic programming and search. *The Journal of Machine Learning Research, 23*(1), 1169–1215.https://dl.acm.org/doi/10.5555/3586589.3586615	especificados para la revisión centrada en mejoras en la explicabilidad de modelos 'caja negra'. La principal razón es que no aborda explícitamente técnicas innovadoras para mejorar la explicabilidad ni proporciona una comparación directa con enfoques tradicionales desde este aspecto. Además, aunque se discuten mejoras en términos de algoritmos y rendimiento, no se evalúa el impacto en la explicabilidad ni se presentan resultados relevantes en este contexto.
21 ACM	Navigating the Ethical Crossroads: Bridging the gap between Predictive Power and Explanation in the use of Artificial Intelligence in Medicine	Riva, G., Sajno, E., De Gaspari, S., Pupillo, C., & Wiederhold, B. K. (Year). Navigating the Ethical Crossroads: Bridging the gap between Predictive Power and Explanation in the use of Artificial Intelligence in Medicine. Annual Review of CyberTherapy and Telemedicine, 21(N/A), 3-7. https://hdl.handle.net/10807/272879	Este artículo explora la dicotomía entre la capacidad predictiva de la inteligencia artificial (IA) en medicina y la habilidad humana para explicar decisiones, destacando los desafíos éticos resultantes. Mientras que los humanos pueden explicar sus elecciones, la IA a menudo opera de manera opaca, generando predicciones sin un razonamiento transparente. Se analizan críticamente las limitaciones de los sistemas de IA médica actuales, resaltando su vulnerabilidad a errores y falta de transparencia. Se argumenta que la explicabilidad es vital para asegurar que los pacientes permanezcan en el centro de la atención médica, capacitándolos para tomar decisiones informadas y autónomas sobre su salud. La Inteligencia Artificial Explicable (XAI) aborda estos desafíos, pero su logro requiere un enfoque multidisciplinario que integre conocimientos tecnológicos con perspectivas psicológicas, cognitivas y sociales. Esta alineación fomentará la innovación, la empatía y la implementación responsable, dando forma a un panorama de atención médica que prioriza tanto el avance tecnológico como las consideraciones éticas.
22 ACM	On the explainability of natural language processing deep models.	El Zini, J., & Awad, M. (2023). On the explainability of natural language processing deep models. ACM Computing Surveys, 55(5), Article 103, 1–31. https://doi.org/10.1145/352975	Cumple con varios de los criterios de inclusión, como la identificación de nuevas metodologías en el contexto de NLP, la propuesta de una taxonomía para métodos de explicabilidad y la discusión sobre el impacto potencial de las mejoras propuestas.
23 ACM	Phase lag index of visual-memory processing EEG for computer-aided AUD diagnosis.	Janah, N. Z., Permanasari, A. E., & Setiawan, N. A. (2023). Phase lag index of visual-memory processing EEG for computer-aided AUD diagnosis. In Proceedings of the 2023 9th International Conference on Computer Technology Applications (ICCTA '23) (pp. 143–150). https://doi.org/10.1145/3605423.360	Se enfoca en el desarrollo de una nueva característica extraída de las señales de EEG para el diagnóstico del trastorno por consumo de alcohol (AUD). El estudio se centra en la creación de una función discriminante que permite diferenciar entre individuos con y sin AUD basándose en la conectividad funcional del cerebro medida a través de EEG.

		5452	
24 ACM	Representation Learning for Maximization of MI, Nonlinear ICA and Nonlinear Subspaces with Robust Density Ratio Estimation.	Sasaki, H., & Takenouchi, T. (2022). Representation Learning for Maximization of MI, Nonlinear ICA and Nonlinear Subspaces with Robust Density Ratio Estimation. Journal of Machine Learning Research, 23(1), 1-55. Submitted 12/20; Revised 6/22; Published 8/22. https://dl.acm.org/doi/10.5555/3586589.3586820	Este artículo aborda el desafío de comprender las representaciones aprendidas en el aprendizaje contrastivo, una técnica del aprendizaje no supervisado. Primero, establece una conexión teórica entre el aprendizaje contrastivo y la maximización de la información mutua (MI), mostrando que la estimación de la razón de densidad es crucial para maximizar la MI. Luego, propone nuevas condiciones de recuperación para componentes de fuentes latentes en el análisis de componentes independientes (ICA) no lineales. Además, presenta un marco novedoso para estimar un subespacio no lineal para componentes de fuentes latentes de menor dimensionalidad, estableciendo condiciones teóricas con la estimación de la razón de densidad. Basándose en estos resultados, propone un método práctico a través de la estimación de la razón de densidad resistente a atípicos, que puede interpretarse como la maximización de MI, ICA no lineal o estimación de subespacio no lineal. Por último, investiga teóricamente la robustez atípica de los métodos propuestos y demuestra su utilidad en experimentos numéricos y una tarea de clasificación lineal.
25 ACM	Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work. In Artificial Intelligence in HCI	Theis, S., Jentzsch, S., Deligiannaki, F., et al. (2023). Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work. In Artificial Intelligence in HCI (Vol. 14050). ISBN: 978-3-031-35890-6.https://doi.org/10.1007/978-3-031-35891-3_22	El paper examina los requisitos para la explicabilidad y aceptación de la inteligencia artificial (IA) en contextos críticos, como el control del tráfico aéreo. Se identifican dos grupos de usuarios: los desarrolladores y los usuarios finales, cuyas necesidades de información varían en especificidad y complejidad. La aceptación de los sistemas de IA depende de varios factores, incluida la función y el rendimiento del sistema, así como consideraciones éticas y de privacidad. Se destaca la importancia de proporcionar información sobre las limitaciones y posibles fallas del sistema. Se sugiere que estos hallazgos pueden orientar futuras investigaciones sobre las necesidades de los usuarios en aplicaciones específicas de IA centradas en el ser humano.
26 ACM	RoCourseNet: Robust Training of a Prediction Aware Recourse Model	Guo, H., Jia, F., Chen, J., Squicciarini, A., & Yadav, A. (2023). RoCourseNet: Robust Training of a Prediction Aware Recourse Model. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (CIKM '23), 619–628. https://doi.org/10.1145/3583780.3615040	El método de explicabilidad utilizado en este caso es la generación de explicaciones contrafácticas (CF). Las explicaciones CF proporcionan casos de recursos contrastivos que muestran qué cambios en las características de entrada podrían haber resultado en una predicción diferente del modelo de aprendizaje automático. En este contexto, RoCourseNet propone un enfoque para generar recuros robustos, es decir, recuros que son válidos incluso cuando el modelo de ML enfrenta

			cambios en la distribución de los datos.
27 ACM	SAMCNet: Towards a Spatially Explainable AI Approach for Classifying MxIF Oncology Data	Farhadloo, M., Molnar, C., Luo, G., Li, Y., Shekhar, S., Maus, R. L., Markovic, S., Leontovich, A., & Moore, R. (2022). SAMCNet: Towards a Spatially Explainable AI Approach for Classifying MxIF Oncology Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD '22), (pp. 2860–2870). https://doi.org/10.1145/3534678.3539168	El abstract describe un enfoque de clasificación de inteligencia artificial (IA) que se centra en la explicabilidad espacial, es decir, en comprender cómo se organizan espacialmente los datos para distinguir entre dos clases (por ejemplo, respondedores y no respondedores). Este enfoque se aplica a datos de puntos multiclase y tiene aplicaciones en la investigación biomédica, específicamente en la búsqueda de nuevas terapias para el cáncer y en la ecología microbiana. El método de explicabilidad utilizado es : Spatially explainable artificial intelligence (IA) classification approach". (SAMCNet)
28 ACM	Stylometric and Semantic Analysis of Demographically Diverse Non-native English Review Data.	Sazzed, S. (2022). Stylometric and Semantic Analysis of Demographically Diverse Non-native English Review Data. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 470-476). Istanbul, Turkey. https://doi.org/10.1109/ASONAM55673.2022.10068612	Se explora cómo diferentes atributos del texto en inglés no nativo varían entre grupos demográficamente distintos. Se utiliza un corpus de alrededor de 1150 reseñas que representan cuatro grupos específicos de países demográficamente diversos: Finlandia, Kenia, Bangladesh y China. Se realizó un análisis estilométrico y semántico en estas reseñas para comprender cómo difieren las características lingüísticas entre las diferentes demografías. Se descubrió que, aunque las características estilométricas son en su mayoría similares entre las reseñas de varios grupos, hay diferencias en atributos como la longitud de la reseña, la presencia de artículos o preposiciones. Se utilizaron algoritmos de aprendizaje automático (ML) clásicos y modelos de lenguaje pre-entrenados basados en transformadores para categorizar las reseñas en grupos demográficos distintos. Se observó que las características semánticas tienen una eficacia ligeramente mejor que las características sintácticas para distinguir las reseñas específicas de cada demografía.
29 ACM	Subgoal-based explanations for unreliable intelligent decision support systems.	Das, D., Kim, B., & Chernova, S. (2023). Subgoal-based explanations for unreliable intelligent decision support systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)* (pp. 240–250). https://doi.org/10.1145/3581641.3584055	Este estudio aborda el desafío de mejorar la interpretabilidad de los sistemas de soporte de decisiones inteligentes (IDS), que a menudo pueden producir resultados subóptimos o fallar en situaciones complejas del mundo real. Se introduce un nuevo tipo de explicación, basada en submetas, para sistemas IDS basados en planes, que complementa la salida tradicional del IDS con información sobre la submeta hacia la cual contribuiría la acción recomendada. Se demuestra que estas explicaciones mejoran el rendimiento de los usuarios en tareas en presencia de recomendaciones del IDS, su capacidad para distinguir entre recomendaciones óptimas y subóptimas, y

			son preferidas por los usuarios. Además, se muestra que las explicaciones basadas en submetas mejoran el rendimiento de los usuarios en caso de falla del IDS, lo que destaca su beneficio significativo en la formación de usuarios para una tarea subyacente.
30 ACM	Techniques for Privacy-Preserving Data Aggregation in an Untrusted Distributed Environment	Shahani, S., Abraham, J., & Venkateswaran. (Year). Techniques for Privacy-Preserving Data Aggregation in an Untrusted Distributed Environment. In Proceedings of the 6th Joint International Conference on Data Science & Management of Data (CODS-COMAD '23) (pp. 286–287). https://doi.org/10.1145/3570991.3571020	El concepto de Privacidad Diferencial (DP) propone un sistema para compartir información sobre un grupo de individuos en un conjunto de datos y proteger su privacidad al mismo tiempo. Caracterizado por parámetros (ϵ, δ), el DP proporciona garantías matemáticas sobre la privacidad y, por lo tanto, es un enfoque reconocido para el análisis y la agregación de datos preservando la privacidad. Los mecanismos de transformación de datos de DP añaden aleatoriedad a los datos para lograr la privacidad diferencial, lo que afecta negativamente la utilidad de los datos, generando un equilibrio entre utilidad y privacidad.
31 ACM	Transparent single-cell set classification with kernel mean embeddings.	Shan, S., Baskaran, V. A., Yi, H., Ranek, J., Stanley, N., & Oliva, J. B. (2022). Transparent single-cell set classification with kernel mean embeddings. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22)* (pp. 1–10). https://doi.org/10.1145/3535508.3545538	Este estudio propone un enfoque para codificar la diversidad celular en muestras biológicas perfiladas utilizando la incrustación media del kernel. Aunque el objetivo principal es mejorar la transparencia del modelo, se observa que este método logra una precisión comparable o superior a los métodos existentes, con menos parámetros. Esto facilita la interpretación de los resultados de clasificación, lo que permite establecer vínculos entre la variabilidad celular y los fenotipos clínicos.
32 ACM	Unlocking the potential of unstructured data in business documents through document intelligence.	Bhatt, H. S., Ramakrishnan, S., Raja, S., & Jawahar, C. V. (2024). Unlocking the potential of unstructured data in business documents through document intelligence. En *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)* (pp. 505–509). https://doi.org/10.1145/3632410.3633293	El abstract se centra en desbloquear el potencial de los documentos no estructurados en el dominio financiero.
33 ACM	Unraveling the impact of explainability of artificial	Ren, Z., Qin, X., & Wang, B. (2023). Unraveling the impact of explainability of artificial intelligence-generated content	El abstract explora cómo la explicabilidad influye en la satisfacción del usuario y las métricas de visión por computadora en herramientas de contenido generado por inteligencia artificial (AIGC) para diseño creativo en

	intelligence-generated content (AIGC) on design style transfer effects	(AIGC) on design style transfer effects. En *Proceedings of the 2023 9th International Conference on Communication and Information Processing (ICCIP '23)* (pp. 171–185). https://doi.org/10.1145/3638884.3638910	interfaces móviles iOS. Examina cuatro herramientas de AIGC y su nivel de explicabilidad, demostrando que una alta explicabilidad conlleva una mayor satisfacción del usuario y evaluación de visión por computadora. Herramientas con niveles medios y bajos de explicabilidad muestran resultados mixtos en satisfacción del usuario y evaluación de visión por computadora. No compara las nuevas metodologías con métodos tradicionales de diseño, se enfoca en cómo los sistemas AIGC cambian la dinámica de diseño tradicional y la integración de la inteligencia artificial en el proceso creativo.
34 ACM	XAI for Medicine by ChatGPT Code interpreter	Kitamura, K., Irvan, M., & Shigetomi Yamaguchi, R. (2023). XAI for Medicine by ChatGPT Code interpreter. In *Proceedings of the 2023 5th International Conference on Big-data Service and Intelligent Computation* (pp. 28–34). https://doi.org/10.1145/3633624.3633629	El abstract presenta un enfoque para mejorar la interpretabilidad de los modelos de lenguaje artificial en el ámbito médico. Se propone el uso de un método llamado Code Base Prompt (CBP) junto con un sistema de evaluación de explicabilidad llamado Medical Algorithm Presentation Criteria (MAPC). CBP consiste en reescribir el algoritmo de toma de decisiones médicas como código Python, mientras que MAPC evalúa la explicabilidad del algoritmo aplicado a texto médico. Se realizó un experimento comparativo utilizando CBP y otro método llamado Text Base Prompt (TBP) en casos de informes médicos sobre insuficiencia cardíaca. Los resultados muestran que CBP logró ejecutar el código Python y cumplir con los criterios de MAPC en todos los casos, mientras que TBP no ejecutó código Python en ninguno de los casos y solo cumplió con un factor de MAPC. Este estudio presenta un nuevo método para implementar la explicabilidad de la inteligencia artificial en tareas médicas utilizando ChatGPT.
35 IEEE	Approaching explainable artificial intelligence methods in the diagnosis of iron deficiency anemia using blood parameters	Ponnusamy, U., D. D. B. S., & Sampathila, N. (2023). Approaching explainable artificial intelligence methods in the diagnosis of iron deficiency anemia using blood parameters. In *Proceedings of the 2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)* (pp. 201-206). IEEE. https://doi.org/10.1109/ICRAIS59684.2023.10367126	El artículo propone un método innovador para diagnosticar anemia ferropénica utilizando parámetros sanguíneos mediante técnicas de machine learning y herramientas de inteligencia artificial explicable (XAI). Estas herramientas, como SHAP y el beeswarm plot, se utilizan para explicar la influencia de los distintos atributos en la clasificación de anemia, mejorando la transparencia y la confianza en los modelos. Los resultados muestran una alta precisión (80-100%) y el método facilita un diagnóstico rápido, ahorrando tiempo a los profesionales de la salud. La colaboración con trabajadores sanitarios y la mejora tecnológica podrían llevar el ámbito médico a nuevos niveles.
36 IEEE	Interpreting black-box machine learning models for high dimensional datasets.	Karim, M. R., et al. (2023). Interpreting black-box machine learning models for high dimensional datasets. In *Proceedings of the 2023 IEEE 10th International Conference on Data Science and*	El artículo presenta un método innovador para mejorar la interpretabilidad de modelos de machine learning 'caja negra' aplicados a datasets de alta dimensionalidad. Primero, se entrena un modelo 'caja negra' en el espacio completo de características para aprender embeddings útiles para la clasificación. Luego, se aplican técnicas de

		Advanced Analytics (DSAA) (pp. 1-10). IEEE. https://doi.org/10.1109/DSAA60987.2023.10302562	probing y perturbación para identificar las características más importantes (explicabilidad global). Un modelo interpretable se entrena sobre estas características seleccionadas, y se derivan reglas de decisión y contra-factuales para proporcionar decisiones locales. Este enfoque supera a métodos como TabNet, XGBoost y técnicas basadas en SHAP en varios datasets con alta dimensionalidad.
37 IEEE	xAI: An Explainable AI Model for the Diagnosis of COPD from CXR Images	Ikechukwu, A.V., & Murali, S. (2023). xAI: An Explainable AI Model for the Diagnosis of COPD from CXR Images. 2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS), 1-6. **DOI:** 10.1109/ICDDS59137.2023.10434619	El estudio aborda la detección temprana de la Enfermedad Pulmonar Obstructiva Crónica (EPOC) utilizando imágenes de radiografías de tórax (CXR) mediante el empleo de algoritmos de deep learning. Se utiliza un conjunto de datos extenso para el pre entrenamiento del modelo y otro conjunto para su desarrollo y validación. Se observa que el modelo basado en Xception, mediante la técnica de fine-tuning, supera al modelo basado en ResNet50 en términos de sensibilidad. Además, se utiliza Grad-CAM y SHAP para proporcionar explicaciones sobre las decisiones del modelo. Los hallazgos resaltan el potencial de los modelos de deep learning en la detección temprana de EPOC a través de CXRs, especialmente en áreas donde las pruebas de espirometría son menos accesibles, y se sugiere la investigación adicional en diagnósticos multimodales de EPOC.
38 IEEE	Planet Optimization with Machine Learning Enabled Power Usage Forecasting Modeling in Smart Grid Environment	Z. a. Almoussawi, W. H. M. Kurdi, B. M. Khaleel, K. AL-Attabi, H. A. Sabah and W. K. Alazzai, "Planet Optimization with Machine Learning Enabled Power Usage Forecasting Modeling in Smart Grid Environment," 2023 6th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, 2023, pp. 726-732,10.1109/IICETA57613.2023.10351470	El abstract cumple parcialmente con los criterios de inclusión establecidos. Identifica una nueva metodología utilizando TWSVM para mejorar la explicabilidad y el rendimiento de la predicción en modelos de machine learning en entornos SG.Carece de una comparación explícita con métodos tradicionales y detalles específicos sobre los resultados cuantitativos. Resumen: El abstract presenta una nueva técnica llamada Planet Optimization with Machine Learning Enabled Power Usage Forecasting Modeling (POML-PUFM) para predecir el consumo de energía eléctrica en entornos de Smart Grids (SG). Se destaca la importancia de la precisión en la predicción del consumo de energía, especialmente en SGs, donde la disponibilidad de datos y la eficiencia del sistema son críticas. La técnica utiliza algoritmos de aprendizaje automático, como Twin-Support Vector Machine (TWSVM), y se basa en la optimización de parámetros para mejorar la precisión de la predicción. Se comparan los resultados obtenidos con otros modelos de aprendizaje automático, destacando el rendimiento superior de POML-PUFM. Los resultados muestran mejoras significativas en la precisión de la predicción del consumo de energía.
39 IEEE	Explainable AI for CPS-Based	Soon, R., Sang, D.V., Chng, C., & Chui, C.K. (2023). Explainable AI	El abstract sobre "Explainable AI for CPS-Based Manufacturing Workcell" tiene puntos positivos, como la

	Manufacturing Workcell	for CPS-Based Manufacturing Workcell. 2023 International Conference on System Science and Engineering (ICSSE), 332-337 . doi: 10.1109/ICSSE58758.2023.10227195	introducción del tema de XAI en entornos manufacturados basados en CPS. Sin embargo, no cumple completamente con los criterios de inclusión establecidos. Específicamente, no se mencionan nuevas metodologías innovadoras para mejorar la explicabilidad, ni se proporciona una comparación clara con métodos tradicionales. Además, faltan detalles sobre el impacto de las mejoras propuestas y resultados prácticos relevantes.
40 IEEE	A Systematic Literature Review of XAI-based Approaches on Brain Disease Detection using Brain MRI Images	Madapatha, S., & Fernando, P. (2024). A Systematic Literature Review of XAI-based Approaches on Brain Disease Detection using Brain MRI Images. In 2024 4th International Conference on Advanced Research in Computing (ICARC) (pp. 19-24). Belihuloya, Sri Lanka. doi: 10.1109/ICARC61713.2024.10499752	Aunque el abstract aborda el uso de XAI como una metodología innovadora para mejorar la explicabilidad de los modelos de IA en el diagnóstico de enfermedades cerebrales por MRI, no cumple completamente con todos los criterios de inclusión establecidos. Específicamente, carece de una comparación explícita con métodos tradicionales y no presenta resultados prácticos que demuestren el impacto de las mejoras en la explicabilidad.
42 IEEE	LCNN: Lightweight CNN Architecture for Software Defect Feature Identification Using Explainable AI	Begum, M., Alam, M., Islam, M. R., & Hossain, M. A. (2024). LCNN: Lightweight CNN Architecture for Software Defect Feature Identification Using Explainable AI. *IEEE Access*, 12, 55744-55756. DOI 10.1109/ACCESS.2024.3388489	El abstract presenta un enfoque para la identificación de defectos de software utilizando inteligencia artificial explicativa (XAI). Se destaca la importancia de mejorar la transparencia y la interpretabilidad en los modelos de inteligencia artificial utilizados para esta tarea. Se describe el uso de dos variantes de redes neuronales convolucionales (CNN) denominadas LCNN, junto con técnicas de preprocesamiento de datos como SMOTE. Se comparan los resultados obtenidos por estas técnicas en términos de precisión, error cuadrático medio (MSE) y área bajo la curva (AUC). Además, se evalúan técnicas XAI como LIME y SHAP para explicar las características identificadas por el modelo. Se concluye que LCNN, especialmente en su versión 2D, supera al 1D-CNN en la identificación de defectos de software. Finalmente, se destaca la eficacia de LIME en la visualización de características de defectos de software, proporcionando una comprensión más clara de las causas subyacentes de los defectos.
43 IEEE	sMRI-PatchNet: A Novel Efficient Explainable Patch-Based Deep Learning Network for Alzheimer's Disease Diagnosis With Structural MRI	Zhang, X., Han, L., Han, L., Chen, H., Dancey, D., & Zhang, D. (2023). sMRI-PatchNet: A Novel Efficient Explainable Patch-Based Deep Learning Network for Alzheimer's Disease Diagnosis With Structural MRI. *IEEE Access, 11,* 108603-108616. https://doi.org/10.1109/ACCESS.2023.3321220 .	El abstract presenta un nuevo enfoque llamado sMRI-PatchNet para el diagnóstico de la enfermedad de Alzheimer utilizando imágenes de resonancia magnética estructural (sMRI). Este método se basa en una red neuronal profunda que utiliza un enfoque de parches para el análisis de imágenes. Propone un método eficiente y explicativo para seleccionar los parches más relevantes del cerebro, lo que mejora la interpretabilidad de los resultados del modelo. Los experimentos muestran que este enfoque puede identificar ubicaciones patológicas de manera efectiva y mejorar el rendimiento en términos de precisión y generalización en comparación con métodos

			existentes.
44 IEEE	Comparing Automated Machine Learning Against an Off-the-Shelf Pattern-Based Classifier in a Class Imbalance Problem: Predicting University Dropou	Cañete-Sifuentes, L., Robles, V., Menasalvas, E., & Monroy, R. (2023). Comparing Automated Machine Learning Against an Off-the-Shelf Pattern-Based Classifier in a Class Imbalance Problem: Predicting University Dropout. IEEE Access, 11, 139147-139156. doi:10.1109/ACCESS.2023.3336596	Se compara el rendimiento de clasificadores automáticos de aprendizaje automático (AutoML) con un clasificador "off-the-shelf" diseñado específicamente para problemas de desequilibrio de clases, utilizando un caso de estudio de predicción de abandono universitario. Se destaca que, aunque AutoML supera a varios clasificadores estándar, el clasificador off-the-shelf logra resultados similares sin la necesidad de costosas etapas de selección y ajuste. Se sugiere la necesidad de que los practicantes de ciencia de datos desarrollen una taxonomía de mecanismos de clasificación y que las plataformas AutoML permitan la modificación del conjunto de clasificadores disponibles y proporcionan explicaciones sobre la selección y el ajuste de los mecanismos.
45 IEEE	Deep Prototypical-Parts Ease Morphological Kidney Stone Identification and are Competitively Robust to Photometric Perturbations.	Flores-Araiza, D., et al. (2023). Deep Prototypical-Parts Ease Morphological Kidney Stone Identification and are Competitively Robust to Photometric Perturbations. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 295-304). Vancouver, BC, Canada. doi: 10.1109/CVPRW59228.2023.00035 .	Se presenta un enfoque para identificar tipos de cálculos renales utilizando técnicas de aprendizaje profundo (DL) que priorizan la explicabilidad. Se propone el uso de Prototypical Parts (PPs) para generar explicaciones interpretables de las decisiones del modelo. Se muestra que este enfoque, aunque tiene una precisión promedio ligeramente inferior a los modelos DL no interpretables, es más robusto frente a perturbaciones en las imágenes. Esto sugiere que aprender PPs puede mejorar la robustez de los modelos DL.
46 IEEE	slidSHAPs – sliding Shapley Values for correlation-based change detection in time series	Balestra, C., Li, B., & Müller, E. (2023). slidSHAPs – sliding Shapley Values for correlation-based change detection in time series. In *Proceedings of the 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, Thessaloniki, Greece (pp. 1-10). IEEE. https://doi.org/10.1109/DSAA60987.2023.10302636	Se presenta slidSHAPs, una metodología para detectar cambios en series temporales multivariadas de manera no supervisada. Se centra en la detección de cambios en la estructura de correlación de los datos de entrada, sin necesidad de etiquetas de clase. slidSHAPs utiliza valores de Shapley para extraer información sobre la estructura de correlación de las secuencias de datos, lo que lo hace más sensible a los cambios que los métodos de detección de puntos de cambio anteriores. Esta técnica ofrece una forma innovadora de mejorar la comprensión y la sensibilidad en la detección de cambios en series temporales.
47 IEEE	Deep Learning for Radar Waveform Design: Retrospectives and the Road Ahead	. Kang, B., Kweon, J., Rangaswamy, M., & Monga, V. (2023). Deep Learning for Radar Waveform Design: Retrospectives and the Road Ahead. IEEE International Radar Conference (RADAR), 1-6. Sydney, Australia. doi: 10.1109/RADAR54928.2023.10371126 .	Se aborda el tema del diseño de formas de onda de radar adaptativas y destaca cómo el aprendizaje profundo se está utilizando cada vez más para abordar este desafío. Se revisan enfoques basados en redes de regresión profunda, como capas completamente conectadas y redes residuales, para abordar problemas de diseño de formas de onda de radar. Se señala la importancia de abordar los problemas de explicabilidad y generalización en estos métodos para garantizar su confiabilidad y practicidad en la implementación.

48 IEEE	Hybrid Intelligence-Driven Medical Image Recognition for Remote Patient Diagnosis in Internet of Medical Things	Guo, Z., Shen, Y., Wan, S., Shang, W. L., & Yu, K. (2022). Hybrid Intelligence-Driven Medical Image Recognition for Remote Patient Diagnosis in Internet of Medical Things. *IEEE Journal of Biomedical and Health Informatics*, 26(12), 5817-5828. doi:10.1109/JBHI.2021.3139541 .	Se presenta un enfoque híbrido de reconocimiento de imágenes médicas para el diagnóstico remoto de pacientes en el Internet de las Cosas Médicas (IoMT). Combina técnicas de aprendizaje profundo y aprendizaje convencional para abordar la falta de explicabilidad en los modelos de aprendizaje profundo. Utiliza una red neuronal convolucional para extraer características profundas de las imágenes y técnicas de aprendizaje convencional para reducir dimensiones y construir un clasificador robusto. Se prueba en un conjunto de datos de miopía patológica, mostrando mejoras en la precisión del reconocimiento.
49 Google Scholar	A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In Proceedings of the 2022	Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) (pp. 2239–2250). https://doi.org/10.1145/3531146.3534639	El abstract aborda temas relacionados con la organización y comprensión del campo de XAI, no proporciona información específica sobre avances recientes en la mejora de la explicabilidad de modelos 'caja negra' ni sobre sus impactos en términos de propuestas, taxonomías u otros resultados relevantes.
50 Google Scholar	A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks.	Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences, 12*(3), 1353. https://doi.org/10.3390/app12031353	El abstract revisa el crecimiento de la explicabilidad en inteligencia artificial (XAI) en respuesta a la necesidad de sistemas más comprensibles. Se destaca la escasez de estudios secundarios en dominios de aplicación específicos y se presenta una revisión sistemática de 137 artículos recientes sobre el desarrollo de métodos de XAI y métricas de evaluación en diversos dominios y tareas. Se concluye que los métodos de XAI se centran principalmente en dominios críticos para la seguridad, como la salud, y que hay una necesidad de más atención en la generación de explicaciones para usuarios generales en dominios sensibles como las finanzas y el sistema judicial. El abstract proporciona una visión general sobre la investigación reciente en XAI y su aplicación en diferentes dominios, pero no cumple completamente con todos los criterios definidos en la pregunta de investigación.
51 Google Scholar	A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts	Schwalbe, G., Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Min Knowl Disc* (2023). https://doi.org/10.1007/s10618-022-00867-8	Se proporciona una visión general de la creación de una taxonomía unificada de métodos de XAI, lo cual es relevante para entender la estructura y clasificación de estos métodos. Sin embargo, no cumple con todos los criterios establecidos para una revisión sistemática sobre XAI, especialmente en términos de identificar nuevas metodologías innovadoras, comparar explícitamente con enfoques anteriores y proporcionar resultados prácticos que demuestren el impacto de las mejoras en la explicabilidad.
52	A comprehensive	Schwalbe, G., & Finzel, B. (2023).	Se aborda el crecimiento exponencial de métodos en el

Google Scholar	taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts	A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*. Advance online publication. https://doi.org/10.1007/s10618-022-00867-8	campo de la inteligencia artificial explicativa (XAI) y la necesidad de una taxonomía completa para comprender, comparar y seleccionar adecuadamente los métodos de XAI. Se destaca que, aunque existen varias taxonomías en la literatura, muchas tienen enfoques diferentes pero comparten puntos de superposición. El estudio presenta una taxonomía unificada basada en una revisión estructurada de más de 50 encuestas sobre métodos de XAI, fusionando conceptos y terminologías de estas encuestas en una taxonomía estructurada. Se resalta que esta taxonomía proporciona una referencia amplia y detallada para investigadores y profesionales, sentando las bases para futuras investigaciones orientadas a casos de uso específicos y contextos sensibles.
53 Google Scholar	Explainable artificial intelligence: A comprehensive review	Minh, D., Wang, H. X., Li, Y. F., et al. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review, 55*, 3503–3568. https://doi.org/10.1007/s10462-021-10088-y	El abstract revisa el crecimiento significativo de la inteligencia artificial explicativa (XAI) debido al aumento en la capacidad de computación y datos. Destaca la importancia de la transparencia en los modelos de IA y cómo la XAI aborda esta falta de claridad al permitir que los algoritmos expliquen sus decisiones internas. El estudio proporciona una revisión detallada de varios métodos de XAI, agrupados en categorías como pre-modelado, modelos interpretables y explicabilidad post-modelado. También analiza los desafíos asociados, como el equilibrio entre el rendimiento y la explicabilidad, y propone enfoques estándar para abordar estos desafíos. En resumen, el abstract presenta una visión general completa de la XAI y su importancia en la comprensión de los modelos de inteligencia artificial.
54 Google Scholar	Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review.	Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon. Advance online publication. DOI:10.1016/j.heliyon.2023.e16110	Se proporciona una visión general de la situación actual y las necesidades futuras en términos de XAI en salud, enfocándose en la evaluación de las propiedades esenciales de XAI y la efectividad de las explicaciones. Sin embargo, no cumple completamente con todos los criterios establecidos para una revisión sistemática sobre XAI, especialmente en términos de identificar nuevas metodologías innovadoras y presentar casos de uso prácticos que demuestren el impacto de las mejoras en la explicabilidad.
55 Google Scholar	Interpretable machine learning for discovery: Statistical challenges and opportunities.	Allen, G. I., Gan, L., & Zheng, L. (2024). Interpretable machine learning for discovery: Statistical challenges and opportunities. *Annual Review of Statistics and Its Application, 11*, 97-121. https://doi.org/10.1146/annurev-statistics-040120-030919	Se ofrece una visión general de cómo el aprendizaje automático interpretable se utiliza para hacer descubrimientos basados en datos y discute la validación de estos descubrimientos, no cumple con varios de los criterios establecidos para evaluar avances recientes en la explicabilidad de modelos 'caja negra'. Específicamente, carece de una identificación clara de nuevas metodologías innovadoras, comparaciones explícitas con enfoques anteriores, y resultados relevantes que demuestren el impacto práctico de las mejoras en la explicabilidad.

56 Google Scholar	A review on interpretable and explainable artificial intelligence in hydroclimatic applications	Başağaoğlu, H., Chakraborty, D., Do Lago, C., Gutierrez, L., Şahinli, M. A., Giacomoni, M., Furl, C., Mirchi, A., Moriasi, D., & Şengör, S. S. (2022). A review on interpretable and explainable artificial intelligence in hydroclimatic applications. *Water, 14*(8), 1230. https://doi.org/10.3390/w14081230	Se proporciona una visión general sobre cómo los modelos de XAI pueden mejorar la explicabilidad en el contexto de predicciones hidroclimáticas, mencionando técnicas innovadoras como Shapley additive explanations y LIME. Sin embargo, falta una comparación más detallada con enfoques anteriores y la presentación de resultados prácticos específicos. Cumple parcialmente con varios criterios establecidos para evaluar avances recientes en la explicabilidad de modelos 'caja negra', pero no de manera completa en todos los aspectos.
57 Google Scholar	Interpretable machine learning for battery capacities prediction and coating parameters analysis.	Liu, K., Faraji Niri, M., Apachitei, G., Lain, M., Greenwood, D., & Marco, J. (2022). Interpretable machine learning for battery capacities prediction and coating parameters analysis. *Control Engineering Practice, 123*, 105202. https://doi.org/10.1016/j.conengprac.2022.105202	Se proporciona una visión general de un marco de aprendizaje automático interpretable aplicado a la fabricación de baterías, destacando la capacidad para predecir propiedades del producto y explicar interacciones de parámetros de fabricación. Cumple parcialmente con varios criterios establecidos para evaluar avances recientes en la explicabilidad de modelos 'caja negra', pero no de manera completa en todos los aspectos. Resumen: Este artículo presenta un framework de machine learning interpretable para predecir las propiedades de las baterías y analizar los parámetros de fabricación. Este enfoque permite predecir con alta precisión tres tipos de capacidades de batería (celular, gravimétrica y volumétrica) en una etapa temprana del proceso de fabricación. También se identifica y cuantifica cómo las variaciones en masa, espesor y porosidad del recubrimiento afectan estas capacidades. El framework es accesible para ingenieros, ya que no requiere conocimiento específico sobre el mecanismo de fabricación de baterías, y proporciona una comprensión detallada de las interacciones entre los parámetros de fabricación y las propiedades finales de las baterías, beneficiando el control inteligente de la fabricación de baterías.
58 Google Scholar	Interpretable and explainable machine learning: A methods-centric overview with concrete examples	Marčinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. First published: 28 February 2023. Edited by: Mehmed Kantardzic, Associate Editor and Witold Pedrycz, Editor-in-Chief. DOI: https://doi.org/10.1002/widm.1493	Se proporciona una visión general sobre el desarrollo de técnicas y modelos para mejorar la interpretabilidad y explicabilidad de modelos 'caja negra', pero no cumple completamente con todos los criterios establecidos para evaluar avances recientes en este campo. Específica algunos avances y modelos actuales, pero no proporciona comparaciones claras con enfoques anteriores ni detalles sobre el impacto específico de estas mejoras en la práctica o la proposición de nuevas taxonomías. Resumen: El artículo destaca la importancia de la interpretabilidad y explicabilidad en el aprendizaje automático y la estadística, especialmente en campos como la medicina, economía, derecho y ciencias naturales. Aunque no existe una definición precisa y universal de estos conceptos, se han desarrollado muchos modelos y técnicas motivados por estas propiedades en las últimas tres décadas, con un enfoque cada vez mayor en el

			aprendizaje profundo. Se presentan ejemplos concretos de modelos de clasificación basados en reglas, escasos y aditivos, aprendizaje de representaciones interpretables y métodos para explicar modelos de caja negra después del análisis. La discusión resalta la necesidad y relevancia de la interpretabilidad y explicabilidad, así como la distinción entre ellas y los sesgos inductivos detrás de los modelos interpretables y los métodos de explicación presentados.
59 Google Scholar	A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging.	Champendal, M., Müller, H., Prior, J. O., & Sá dos Reis, C. (2023). A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. European Journal of Radiology, 169, 111159 . DOI:https://doi.org/10.1016/j.ejrad.2023.111159	El abstract proporciona una visión general sobre el uso creciente de técnicas de XAI en imágenes médicas, destacando la aplicación en diversas modalidades y patologías. Aunque cumple con algunos criterios de inclusión, como la identificación de nuevas metodologías y la presentación de resultados relevantes, no cumple con todos los criterios, especialmente en términos de comparación con enfoques anteriores, impacto estructurado de las mejoras y propuestas de taxonomías específicas. Resumen:el abstract presenta una revisión de los métodos de Inteligencia Artificial Explicable (XAI) disponibles para la imagen médica. Se realizó una revisión de alcance siguiendo la metodología del Instituto Joanna Briggs, buscando en varias bases de datos y utilizando palabras clave relacionadas con la explicabilidad y las modalidades de imagen médica. Se identificaron 228 estudios que cumplían con los criterios de inclusión, mostrando un aumento en las publicaciones de XAI, especialmente en MRI, radiografía y CT, dirigidas principalmente a la clasificación y predicción de patologías pulmonares y cerebrales. Las explicaciones se presentan principalmente de forma visual y numérica, con enfoques post-hoc predominantemente utilizados. La estandarización de la terminología sigue siendo un desafío. En resumen, el desarrollo futuro de XAI debe considerar las necesidades y perspectivas de los usuarios.
60 Google Scholar	Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey.	Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*. Advance online publication. https://doi.org/10.1016/j.ins.2022.10.013	El abstract proporciona una visión general sobre XAI, destacando su importancia, desafíos y contribuciones potenciales. Aunque cumple con algunos criterios de inclusión, como la identificación de nuevas metodologías y la presentación de resultados relevantes, no cumple con todos los criterios, especialmente en términos de comparación detallada con enfoques anteriores, propuestas nuevas específicas y frameworks que indiquen un impacto estructurado de las mejoras en la explicabilidad. Resumen:El estudio explora el campo de la Inteligencia Artificial Explicable (XAI), que busca desentrañar los modelos de IA considerados caja negra mediante la generación de explicaciones comprensibles. Se presenta una nueva definición de explicabilidad y se propone una

			taxonomía detallada para categorizar los estudios de XAI. Se realizó un análisis comparativo experimental de algoritmos de XAI aplicados a diferentes tipos de datos. Se discuten las implicaciones éticas y prácticas de XAI, así como las preguntas de investigación abiertas y los desafíos futuros.
61 Google Scholar	Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models.	Salih, A., Boscolo Galazzo, I., Gkontra, P., Lee, A. M., Lekadir, K., Raisi-Estabragh, Z., & Petersen, S. E. (2023). Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models. *Circulation: Cardiovascular Imaging, 16*. https://doi.org/10.1161/CIRCIMAGING.122.014519	El abstract proporciona una visión general sobre la necesidad y el potencial de XAI en la imagen cardiaca, pero no cumple con todos los criterios establecidos para evaluar los avances en la explicabilidad de modelos 'caja negra'. Específicamente, carece de una comparación explícita con enfoques anteriores en términos de mejoras en la explicabilidad y de detalles sobre nuevas propuestas o frameworks que indiquen un impacto estructurado de las mejoras en la explicabilidad. Resumen:El abstract proporciona una revisión exhaustiva de la literatura sobre el uso de métodos de Inteligencia Artificial Explicable (XAI) en el campo de la imagen cardíaca. Destaca la necesidad de moverse hacia modelos más interpretables en lugar de depender exclusivamente de modelos de aprendizaje profundo considerados cajas negras. Además, ofrece pautas simples y comprensibles sobre XAI en el contexto de la imagen cardíaca. También aborda los problemas abiertos y las direcciones futuras para la aplicación de XAI en este campo específico.
62 Google Scholar	Explainability and Interpretability in Electric Load Forecasting Using Machine Learning Techniques – A Review	Baur, L., Ditschuneit, K., Schambach, M., Kaymakci, C., Wollmann, T., & Sauer, A. (2024). Explainability and Interpretability in Electric Load Forecasting Using Machine Learning Techniques – A Review. *Energy AI, 1*, 100358. https://doi.org/10.1016/j.egyai.2024.100358	El abstract proporciona una visión general sobre el estado actual y las tendencias en el uso de métodos interpretables y explicables para el pronóstico de carga eléctrica. Cumple con algunos criterios clave al identificar nuevas metodologías y técnicas aplicadas, así como al presentar una taxonomía de enfoques. Sin embargo, no cumple completamente con la comparación explícita con métodos anteriores, ni con la presentación de propuestas nuevas específicas o estudios de caso prácticos que demuestren el impacto de las mejoras en la explicabilidad. Resumen:El abstract revisa el estado actual de los métodos de Machine Learning aplicados a la predicción de la carga eléctrica. Destaca la importancia de que estos modelos sean interpretables y explicables debido a la creciente automatización en la gestión de la demanda de electricidad. Se realiza una revisión de la literatura para identificar enfoques aplicados en la interpretabilidad y explicabilidad de las predicciones de carga eléctrica utilizando Machine Learning. Se observa un aumento en el uso de modelos probabilísticos y técnicas de lógica difusa, así como enfoques como la importancia de características y los mecanismos de atención para explicar las predicciones. Se discuten las tendencias y se señala que aún queda mucho por adaptar de otros campos, como el pronóstico de series temporales, para mejorar la

			interpretabilidad en la predicción de carga eléctrica.
63 Google Scholar	Interpretable artificial intelligence in radiology and radiation oncology.	Cui, S., Traverso, A., Niraula, D., Zou, J., Luo, Y., Owen, D., El Naqa, I., Wei, L. (2023). Interpretable artificial intelligence in radiology and radiation oncology. *British Journal of Radiology*, 96(1150), 20230142. https://doi.org/10.1259/bjr.20230142	Se ofrece una introducción a la importancia de la interpretabilidad en modelos de IA aplicados en radiología y oncología radioterápica, pero no cumple completamente con los criterios establecidos para evaluar avances recientes en la explicabilidad de modelos 'caja negra'. Aunque proporciona una visión general y discute algunos conceptos y métodos, carece de evidencia directa sobre nuevas metodologías, comparaciones con enfoques anteriores, impacto estructurado de mejoras y resultados prácticos relevantes.
64 Google Scholar	A spectrum of explainable and interpretable machine learning approaches for genomic studies	Conard, A. M., DenAdel, A., & Crawford, L. (2023). A spectrum of explainable and interpretable machine learning approaches for genomic studies. *Wiley Interdisciplinary Reviews: Computational Statistics, 15*, e1617. https://doi.org/10.1002/wics.1617	Se destaca la importancia de la transparencia en los modelos de aprendizaje automático aplicados a la genómica. Señala que, aunque los modelos de aprendizaje automático han sido efectivos para predicciones en biomedicina, su naturaleza de "caja negra" dificulta la comprensión de cómo hacen estas predicciones. El artículo revisa métodos explicables e interpretables que buscan abordar esta limitación, enfocándose en aplicaciones genómicas y resaltando la incorporación de conocimientos biológicos en el desarrollo de estos métodos.
65 Google Scholar	Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency.	Barragán-Montero, A., Bibal, A., Dastarac, M., Draguet, C., Valdes, G., Nguyen, D., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Souris, K., Sterpin, E., & Lee, J. (2022). Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. *Physics in Medicine & Biology, 67*. https://doi.org/10.1088/1361-6560/ac678a	El abstract proporciona una visión general de los desafíos de interpretabilidad de modelos ML en oncología radioterápica,. Se revisa el creciente interés en el aprendizaje automático (ML) en radioterapia, destacando su complejidad y la necesidad de interpretabilidad. Señala los riesgos asociados con los datos y los modelos de ML, así como la interacción entre ellos, y destaca la importancia de la interpretabilidad y la dependencia entre datos y modelos. Define formalmente conceptos clave como interpretabilidad, explicabilidad y dependencia entre datos y modelos, y discute aplicaciones de ML en radioterapia y perspectivas de implementación clínica.
66 Google Scholar	Artificial Intelligence–HRM Interactions and Outcomes: A Systematic Review and Causal Configurational Explanation.	Basu, S., Majumdar, B., Mukherjee, K., Munjal, S., & Palaksha, C. (2023). Artificial Intelligence–HRM Interactions and Outcomes: A Systematic Review and Causal Configurational Explanation. *Human Resource Management Review, 33*(1), 100893. https://doi.org/10.1016/j.hrmr.2022.100893	El abstract ofrece una revisión sistemática de la literatura sobre la interacción entre inteligencia artificial (IA) y gestión de recursos humanos (HRM). Destaca cómo los sistemas de IA están siendo integrados en diversas funciones organizativas, señalando tanto sus beneficios para el rendimiento organizacional como las preocupaciones sobre pérdidas de empleo. Identifica configuraciones causales temáticas en la investigación, analizando su evolución y proporcionando explicaciones basadas en estas configuraciones para los resultados en la interacción entre IA y HRM.

67 Google Scholar	Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence–Based Health Technologies: What Healthcare Stakeholders Need to Know	Farah, L., Murris, J. M., Borget, I., Guilloux, A., Martelli, N. M., & Katsahian, S. I. M. (2023). Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: What healthcare stakeholders need to know. *Mayo Clinic Proceedings: Digital Health, 1*(2), 120-138. https://doi.org/10.1016/j.mcpdig.2023.02.004	El abstract destaca la importancia de la interpretabilidad y la explicabilidad en los dispositivos médicos basados en inteligencia artificial (IA). Se realizó una revisión de literatura para identificar los criterios clave necesarios en la evaluación de estos dispositivos, destacando la necesidad de evaluar el rendimiento, la interpretabilidad y la explicabilidad de los algoritmos de IA. Se proporcionan recomendaciones sobre cómo evaluar el rendimiento y se describen formas de apoyar la evaluación de la interpretabilidad y explicabilidad. Además, se propone un conjunto de herramientas y métodos para comprender cómo funcionan los algoritmos de aprendizaje automático y sus predicciones en el contexto de la atención médica.
68 Google Scholar	Interpretable artificial intelligence and exascale molecular dynamics simulations to reveal kinetics: Applications to Alzheimer's disease.	Martin, W., Sheynkman, G., Lightstone, F. C., Nussinov, R., & Cheng, F. (2022). Interpretable artificial intelligence and exascale molecular dynamics simulations to reveal kinetics: Applications to Alzheimer's disease. *Current Opinion in Structural Biology, 72,* 103-113. https://doi.org/10.1016/j.sbi.2021.09.001	El abstract revisa cómo las técnicas de aprendizaje, como la inteligencia artificial, se utilizan en simulaciones de dinámica molecular para comprender la agregación de proteínas en la enfermedad de Alzheimer. Destaca la necesidad de hacer que los datos generados sean interpretables e introduce conceptos de aprendizaje. Aunque aborda la mejora de la explicabilidad de los modelos de inteligencia artificial, no ofrece una comparación explícita con enfoques anteriores ni evalúa el impacto práctico de estas mejoras.
69 Google Scholar	Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience	Goodwin, N. L., Nilsson, S. R. O., Choong, J. J., & Golden, S. A. (2022). Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. *Neurobiology of Behavior 2022*, Edited by Tiago Branco and Mala Murthy. University of Washington.	El abstract destaca el uso de técnicas innovadoras, como los valores de Shapley a través de Shapley Additive Explanations (SHAP), para mejorar la explicabilidad de los modelos de machine learning en el campo de la neuroetología computacional. Se enfoca en tres aplicaciones principales de estas herramientas de explicabilidad: estandarización, especialización y explicabilidad. Además, se sugiere que estas herramientas pueden ayudar a eliminar el sesgo manual y a identificar repertorios de comportamiento previamente desconocidos. Se subraya la importancia de estas herramientas como un paso necesario para avanzar en el campo y propone el uso de valores de Shapley como una posible solución para mejorar la explicabilidad en el análisis de aprendizaje automático.
70 Google Scholar	Comparing Explanation Methods for Traditional Machine Learning Models Part 1: An Overview of Current Methods and Quantifying Their Disagreement.	Flora, M., Potvin, C., McGovern, A., & Handler, S. (2022, November 16). Comparing Explanation Methods for Traditional Machine Learning Models Part 1: An Overview of Current Methods and Quantifying Their Disagreement. DOI:10.48550/arXiv.2211.08943	Se proporciona una visión general y crítica de métodos de explicación para modelos 'caja negra' en ML, cumpliendo con algunos criterios de inclusión como la identificación de nuevas metodologías y la comparación con enfoques anteriores. Sin embargo, no cumple completamente con los criterios de impacto estructurado de mejoras en la explicabilidad y propuestas nuevas o taxonomías específicas.

71 Google Scholar	Interpretability and Explainability of Machine Learning Models: Achievements and Challenges	Henriques, J., Rocha, T., de Carvalho, P., Silva, C., Paredes, S. (2024). Interpretability and Explainability of Machine Learning Models: Achievements and Challenges. En: Pino, E., Magjarević, R., de Carvalho, P. (eds) Conferencia Internacional sobre Informática Biomédica y de Salud 2022. ICBHI 2022. Actas de IFMBE, vol. 108. Springer, Cham. https://doi.org/10.1007/978-3-031-59216-4_9	El abstract destaca la necesidad de mejorar la explicabilidad de los modelos de aprendizaje automático en el ámbito clínico. Se menciona el campo emergente de la Inteligencia Artificial Explicable (XAI) como una solución para este problema, proponiendo enfoques que mantengan altos niveles de rendimiento mientras mejoran la explicabilidad de los modelos. Se sugiere que esto conduciría a una mejor comprensión de cómo funcionan los modelos, aumentaría la confianza en su uso y ayudaría a los profesionales de la salud en la toma de decisiones clínicas.
72 Google Scholar	Interpretable artificial intelligence systems in medical imaging: Review and theoretical framework	Xian, T., Constantinides, P., & Mehandjiev, N. (2024). Interpretable artificial intelligence systems in medical imaging: Review and theoretical framework. En E. Pino, R. Magjarević y P. de Carvalho (Eds.), *Business 2024* (pp. 240–265). DOI: https://doi.org/10.4337/9781803926216.00023	El abstract revisa el impacto de la inteligencia artificial (IA) interpretable en la toma de decisiones médicas, centrándose en el análisis de imágenes médicas. Proporciona un marco de referencia para comprender los sistemas de IA interpretables, destacando tres componentes clave: agentes humanos, datos y modelos de aprendizaje automático (ML). Utilizando el proceso de cribado de cáncer de mama como ejemplo, identifica posibles tensiones entre los agentes humanos y los modelos de ML. El estudio concluye con implicaciones para futuras investigaciones en este campo.
73 Google Scholar	Pediatrics in Artificial Intelligence Era: A Systematic Review on Challenges, Opportunities, and Explainability	Balla, Y., Tirunagari, S., & Windridge, D. (2023). Pediatrics in Artificial Intelligence Era: A Systematic Review on Challenges, Opportunities, and Explainability. *Indian Pediatrics*, 60, 561–569. https://doi.org/10.1007/s13312-023-2936-8	El estudio revisa el papel de la inteligencia artificial (IA) en la medicina pediátrica, destacando los desafíos, oportunidades y la necesidad de explicabilidad de los modelos de IA. Se realizó una búsqueda sistemática de artículos entre 2016 y 2022, identificando 20 estudios relevantes. Tres temas principales surgieron: la aplicación actual de IA en el diagnóstico y predicción de condiciones de salud pediátricas, los desafíos específicos de implementar IA en medicina pediátrica y las oportunidades futuras para adaptar la IA. El estudio enfatiza que la IA debe ser vista como una herramienta para mejorar y respaldar la toma de decisiones clínicas, y sugiere que la investigación futura se centre en obtener datos completos para garantizar la generalización de los hallazgos.
74 Google Scholar	A comprehensive review and application of interpretable deep learning model for ADR prediction	Dubey, S. A., & Pandit, A. A. (2022). A comprehensive review and application of interpretable deep learning model for ADR prediction. *International Journal of Advanced Computer Science and Applications (IJACSA), 13*(9). http://dx.doi.org/10.14569/IJACSA.2022.0130924	Este estudio aborda la seguridad de los medicamentos mediante la detección y predicción de reacciones adversas a los medicamentos (ADR). Se recopilaron 172 artículos de investigación de bases de datos como ResearchGate y PubMed, que se clasificaron en temas de detección y predicción de ADR. Se analizaron fuentes de datos comunes, algoritmos y métricas de evaluación. Se diseñó e implementó un modelo de deep learning con dos capas ocultas, que mostró un rendimiento óptimo para la predicción de ADR. Para mejorar la interpretabilidad del modelo, se utilizó un modelo surrogate global. La

			arquitectura propuesta superó varias limitaciones de los modelos existentes y destacó la importancia de la detección y predicción temprana de ADR en la industria de la salud.
75 Google Scholar	Artificial intelligence and explanation: How, why, and when to explain black boxes	Marcus, E., & Teuwen, J. (2024). Artificial intelligence and explanation: How, why, and when to explain black boxes. *European Journal of Radiology, 173*, 111393. https://doi.org/10.1016/j.ejrad.2024.111393	El artículo aborda el desafío de la explicabilidad en inteligencia artificial (IA), especialmente en el contexto médico como la radiología. Se destaca que los algoritmos de IA a menudo se consideran "cajas negras" y la falta de explicabilidad podría afectar aspectos críticos como la confianza del paciente y la detección de errores. El estudio argumenta que incluso los modelos más complejos pueden entenderse, utilizando analogías con la comprensión de las leyes físicas. Se discute el proceso de explicación y se enfatiza el papel tanto de los desarrolladores de IA como de los profesionales médicos, como los radiólogos, en la mejora continua de los modelos de IA. Además, se explora el papel del programa de IA explicativa (XAI) en este contexto más amplio.
76 Google Scholar	Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work	Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A. P., & Bruder, C. (2023). Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work. En H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI. HCII 2023*. Lecture Notes in Computer Science (Vol. 14050). Springer, Cham. https://doi.org/10.1007/978-3-031-35891-3_22	El análisis estructurado de la literatura examina los requisitos para la explicabilidad y aceptación de la inteligencia artificial (IA). Se revisan artículos sobre la información necesaria para percibir una IA como explicable, la información necesaria para aceptar una IA y los métodos de representación e interacción que promueven la confianza en una IA. Se destaca la importancia de proporcionar información a dos grupos principales de usuarios: los desarrolladores, que requieren información sobre las operaciones internas del modelo, y los usuarios finales, que necesitan información sobre los resultados o el comportamiento de la IA. La aceptación de los sistemas de IA depende de la información sobre las funciones y el rendimiento del sistema, consideraciones de privacidad y ética, así como información para establecer la confianza en el sistema. Se sugiere que la información sobre las limitaciones y posibles fallas del sistema puede aumentar la aceptación y la confianza. Los métodos de interacción confiables son similares a los humanos e incluyen lenguaje natural, habla, texto y representaciones visuales como gráficos, tablas y animaciones. Estos resultados tienen implicaciones significativas para el desarrollo futuro de sistemas de IA centrados en el ser humano.
77 Google Scholar	Explainable and interpretable machine learning for antimicrobial stewardship: Opportunities and challenges.	Giacobbe, D. R., Marelli, C., Guastavino, S., Mora, S., Rosso, N., Signori, A., Campi, C., Giacomini, M., & Bassetti, M. (2024). Explainable and interpretable machine learning for antimicrobial stewardship: Opportunities and challenges. *Clinical Therapeutics*.	Se discute el interés creciente en el uso de inteligencia artificial y machine learning para mejorar la prescripción antimicrobiana, destacando la importancia de la interpretabilidad y explicabilidad para evitar sesgos no deseados. Se revisan algunos temas relevantes sobre el uso de algoritmos de machine learning para intervenciones de gestión antimicrobiana, resaltando oportunidades y desafíos, con un enfoque particular en la interpretabilidad

		Advance online publication. https://doi.org/10.1016/j.clinthera.2024.02.010	y explicabilidad de los modelos empleados. Se sugiere que comprender cómo funcionan los modelos de machine learning puede mejorar la prescripción antimicrobiana y ayudar a prevenir la resistencia antimicrobiana.
78 Google Scholar	A critical moment in machine learning in medicine: on reproducible and interpretable learning	Ciobanu-Caraus, O., Aicher, A., Kernbach, J.M. *et al.* A critical moment in machine learning in medicine: on reproducible and interpretable learning. *Acta Neurochir* 166, 14 (2024). https://doi.org/10.1007/s00701-024-05892-8	Se discute el crecimiento exponencial de las publicaciones sobre machine learning (ML) en los últimos años y cómo la falta de rigor metodológico y pautas de informes estándar ha contribuido a una crisis de reproducibilidad en este campo. Además, destaca cómo la creciente complejidad de los modelos de ML compromete su interpretabilidad, lo que obstaculiza su adopción clínica. Se exploran posibles soluciones para contrarrestar este problema, como el desarrollo de pautas de informes estándar, el fomento de la compartición de datos y código, y el uso de herramientas de explicación y análisis de sensibilidad. Se enfatiza la importancia de equilibrar el rendimiento del modelo con su interpretabilidad para garantizar su aplicabilidad clínica. Este abstract ofrece una visión crítica sobre los desafíos actuales en el campo del ML en medicina y sugiere posibles vías para abordarlos.
79 Google Scholar	The Pros and Cons of Using Machine Learning and Interpretable Machine Learning Methods in psychiatry detection applications, specifically depression disorder: A Brief Review	Smith, J., & Johnson, A. (2023). The Pros and Cons of Using Machine Learning and Interpretable Machine Learning Methods In Psychiatry Detection Applications Specifically Depression Disorder: A Brief Review. ResearchGate. https://www.researchgate.net/publication/375331868/The_Pros_and_Cons_of_Using_Machine_Learning_and_Interpretable_Machine_Learning_Methods_In_Psychiatry_Detection_Applications_Specifically_Depression_Disorder_A_Brief_Review	El artículo destaca el aumento de la importancia del uso de machine learning en el diagnóstico de trastornos psiquiátricos, especialmente la depresión, durante la pandemia de COVID-19. Se enfatiza la necesidad de desarrollar soluciones de inteligencia artificial interpretables para proporcionar diagnósticos precisos y comprensibles. Sin embargo, el resumen no aborda directamente nuevos avances en la explicabilidad de modelos de "caja negra" ni proporciona comparaciones claras entre enfoques anteriores y nuevos métodos en términos de explicabilidad.
80 Google Scholar	A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System	Sheu, R.-K., & Pardeshi, M. S. (2022). A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors*, 22(20), 8068. https://doi.org/10.3390/s22208068	El abstract presenta una revisión detallada sobre el uso de eXplainable AI (XAI) en el ámbito médico, enfocándose en la importancia de comprender las decisiones detalladas, resultados y condiciones de los pacientes. Se discuten avances recientes, métodos de evaluación, estudios de casos y mejoras futuras en XAI para la medicina. Se destacan diferencias potenciales entre IA y XAI, y se propone un enfoque de colaboración humano-máquina para producir soluciones explicables. Además, se introduce un sistema de puntuación y recomendación de XAI.
81 Google	Explainable and interpretable	Frasca, M., La Torre, D., Pravettoni, G., et al. (2024). Explainable and	Se analiza el creciente impacto de los algoritmos de aprendizaje automático y aprendizaje profundo en el

Scholar	artificial intelligence in medicine: A systematic bibliometric review. Discovery of Artificial Intelligence	interpretable artificial intelligence in medicine: A systematic bibliometric review. Discovery of Artificial Intelligence, 4(15), 1-14. https://doi.org/10.1007/s44163-024-00114-7	campo médico, centrándose en los problemas críticos de explicabilidad e interpretabilidad asociados con los algoritmos de caja negra. Se revisan desafíos y soluciones en la literatura, ofreciendo una visión general de las técnicas más recientes utilizadas en este campo y definiciones precisas de estos conceptos. Se destaca el crecimiento exponencial en el campo en la última década, con un énfasis en la necesidad de una comunicación efectiva sobre las capacidades y limitaciones de la inteligencia artificial en la toma de decisiones médicas. Se discuten las dimensiones psicológicas de la percepción pública y se aboga por un compromiso constante con la transparencia, la ética y la colaboración interdisciplinaria.
82 Google Scholar	Assessing XAI: Unveiling Evaluation Metrics forLocal Explanation, Taxonomies, Key Concepts, andPractical Applications	Kadir, M. A., Mosavi, A., & Sonntag, D. (2023, May 5). Assessing XAI: Unveiling Evaluation Metrics for Local Explanation, Taxonomies, Key Concepts, and Practical Applications. German Research Center for Artificial Intelligence & John von Neumann Faculty of Informatics, Obuda University.	Se revisan los avances recientes en la mejora de la explicabilidad de los modelos de machine learning, centrándose en los métodos de inteligencia artificial explicables (XAI). Se destaca la necesidad de definir claramente la validez, confiabilidad y métricas de evaluación de la explicabilidad debido a la diversidad de datos y metodologías de aprendizaje. Utilizando la guía sistemática PRISMA para una revisión exhaustiva de la literatura, se examinan las métricas de evaluación utilizadas para XAI. Basándose en los resultados, se proponen dos taxonomías para estas métricas, una basada en aplicaciones y otra en métricas de evaluación.
83 Google Scholar	CSG compressive strength prediction based on LSTM and interpretable machine learning	Tian, Q., Gao, H., Guo, L., Li, Z., & Wang, Q. (2023). CSG compressive strength prediction based on LSTM and interpretable machine learning. *Reviews on Advanced Materials Science*, 62(1), 20230133. https://doi.org/10.1515/rams-2023-0133	El estudio propone un enfoque para predecir la resistencia a la compresión del cemento utilizando un modelo de aprendizaje profundo LSTM y técnicas de aprendizaje automático interpretable. Se comparan los resultados con un modelo de bosque aleatorio (RF), demostrando que el LSTM tiene una precisión y fiabilidad superiores. Además, se utiliza el método SHapley Additive exPlanations para explicar la contribución de cada característica de entrada en el modelo de aprendizaje automático. Los resultados muestran que el cemento y la tasa de arena son los contribuyentes más significativos a las predicciones.
84 Google Scholar	Interpretable and Explainable Machine Learning Methods for Predictive Process Monitoring: A Systematic Literature Review	Mehdiyev, N., Majlatow, M., & Fettke, P. (2023). Interpretable and Explainable Machine Learning Methods for Predictive Process Monitoring: A Systematic Literature Review. German Research Center for Artificial Intelligence (DFKI), Saarland University. Retrieved from https://arxiv.org/pdf/2312.17584	El abstract proporciona una visión general sobre la revisión sistemática de la literatura en explicabilidad e interpretabilidad de modelos de ML en minería predictiva de procesos. Cumple con varios criterios, incluyendo la identificación de nuevas metodologías y la comparación con enfoques anteriores, pero no proporciona propuestas o taxonomías nuevas específicas. Resumen: El artículo presenta una revisión sistemática de la literatura sobre la explicabilidad e interpretabilidad de modelos de aprendizaje automático en el contexto de la minería predictiva de procesos. Se destaca la importancia de comprender los modelos "caja negra" en este dominio específico. Se diferencia entre modelos intrínsecamente

			interpretables y aquellos que requieren técnicas de explicación posteriores. El estudio ofrece una síntesis detallada de metodologías actuales y su aplicación en diversos dominios. Los hallazgos buscan proporcionar una comprensión más profunda para el desarrollo de sistemas inteligentes más transparentes y efectivos.
85 Google Scholar	Explainability in Mechanism Design: Recent Advances and the Road Ahead.	Alape Suryanarayana, S., Sarne, D., & Kraus, S. (2022). Explainability in Mechanism Design: Recent Advances and the Road Ahead. arXiv preprint arXiv:2206.03031. [https://doi.org/10.48550/arXiv.2206.03031]	El abstract del artículo presenta un análisis exhaustivo sobre la explicabilidad en el diseño de mecanismos, una disciplina que involucra agentes económicamente motivados y decisiones que no necesariamente maximizan todas las funciones de utilidad individuales. Se distingue entre modelos intrínsecamente interpretables y aquellos que requieren técnicas de explicación adicionales. Además, se discuten las propiedades y objetivos principales de la explicabilidad en este contexto, y se proponen algunos conceptos de solución para abordar los desafíos identificados. Aunque no se enfoca específicamente en modelos de aprendizaje automático considerados "caja negra", el artículo proporciona una visión general valiosa de la explicabilidad en un contexto diferente, ofreciendo una contribución significativa al campo de los sistemas inteligentes y la toma de decisiones.
86 Google Scholar	A Taxonomy of Non-Fungible Tokens: Overview, Evaluation and Explanation	Olsson, O. (2022). A Taxonomy of Non-Fungible Tokens: Overview, Evaluation and Explanation (Ph.D. Dissertation). Uppsala Universitet, Sweden. Order Number: AAI29353266.	El resumen presenta una visión general del papel fundamental que desempeñan los tokens no fungibles (NFT) en el contexto de la web 3.0 y la tecnología blockchain. Destaca cómo los NFTs están transformando la representación de la propiedad digital al permitir la transferencia de activos digitales de manera única y segura a través de contratos inteligentes en la cadena de bloques. La revisión sistemática de la literatura sobre NFTs ha dado lugar a una taxonomía inicial que desglosa los componentes esenciales de estos tokens. Además, se ha evaluado y ampliado esta taxonomía para reflejar mejor los proyectos reales relacionados con NFTs. En última instancia, el estudio contribuye a la comprensión y la organización del campo emergente de investigación sobre NFTs mediante la presentación de una estructura organizativa y descriptiva.
87 Google Scholar	Tensor Networks for Interpretable and Efficient Quantum-Inspired Machine Learning	Ran, S.-J., & Su, G. (2023, November 17). Tensor Networks for Interpretable and Efficient Quantum-Inspired Machine Learning. *Intelligent Computing, 2*, Article ID: 0061. https://doi.org/10.34133/icomputing.0061	El abstract presenta una revisión breve pero informativa sobre el uso de redes tensoriales (TN) en el aprendizaje automático (ML). Destaca el desafío actual de lograr alta interpretabilidad y eficiencia en los modelos de ML, y cómo las TN, inspiradas en la mecánica cuántica, ofrecen una solución prometedora. Se menciona la solidez teórica de las TN en términos de información cuántica y física de sistemas de muchos cuerpos, así como su eficiencia computacional. Además, se sugiere que las TN podrían ser una herramienta fundamental para desarrollar esquemas de "inteligencia artificial cuántica" en el futuro.

88 Google Scholar	Advancements in Deep Reinforcement Learning and Inverse Reinforcement Learning for Robotic Manipulation: Toward Trustworthy, Interpretable, and Explainable Artificial Intelligence	Ozalp, R., Ucar, A., & Guzelis, C. (2024). Advancements in Deep Reinforcement Learning and Inverse Reinforcement Learning for Robotic Manipulation: Toward Trustworthy, Interpretable, and Explainable Artificial Intelligence. *IEEE Access, 12*, 51840-51858. https://doi.org/10.1109/ACCESS.2024.3385426	El artículo revisa los avances en el uso de Aprendizaje Profundo por Refuerzo (DRL) y Aprendizaje por Refuerzo Inverso (IRL) en tareas de manipulación robótica durante los últimos cinco años. Se examinan varios aspectos, como la percepción, el ensamblaje, la manipulación con recompensas inciertas, el multitasking, el aprendizaje de transferencia, la multimodalidad y la interacción humano-robot (HRI). Se resumen los principales aportes, métodos y desafíos de los estudios relevantes, junto con tablas de resumen sobre el problema y la solución. Además, se discuten los conceptos de IA confiable, IA interpretable y IA explicable (XAI) en el contexto de la manipulación robótica, proporcionando así un recurso para futuras investigaciones en este campo.
89 Google Scholar	Targets of explanation in correctional and forensic psychology: A black box model	Ward, T., & Durrant, R. (2022). Targets of explanation in correctional and forensic psychology: A black box model. *Aggression and Violent Behavior, 67*, 101782. https://doi.org/10.1016/j.avb.2022.101782	El abstract aborda la falta de claridad en los objetivos de las teorías explicativas en psicología correccional, señalando un exceso de tiempo dedicado a construir explicaciones mal dirigidas. Propone pasos para abordar este problema, incluyendo soluciones centradas en problemas psicológicos y sociales asociados con la conducta delictiva, así como una solución basada en el concepto de "cajas negras". Aunque no menciona nuevas metodologías o comparaciones con enfoques anteriores, ofrece una perspectiva crítica sobre cómo enfocar la explicabilidad en este ámbito. Sin embargo, no proporciona resultados prácticos específicos o taxonomías que indiquen el impacto de las mejoras en la explicabilidad. En general, ofrece una visión sobre cómo abordar el problema de la explicabilidad en modelos "caja negra", pero carece de algunos aspectos clave.
90 Google Scholar	Catching Silent Failures: A Machine Learning Model Monitoring and Explainability Survey	Karval, R., & Singh, K. N. (2023). Catching Silent Failures: A Machine Learning Model Monitoring and Explainability Survey. In 2023 OITS International Conference on Information Technology (OCIT) (pp. 526-532). Raipur, India. doi:10.1109/OCIT59427.2023.10431343 .	El artículo aborda la importancia de la monitorización y explicabilidad de los modelos de aprendizaje automático (ML) una vez que son liberados para su uso. Se plantea la pregunta de si los resultados producidos por estos modelos cumplen con las necesidades de los usuarios finales y contribuyen a la construcción de mejores sistemas. Se destaca la necesidad de establecer confianza, responsabilidad y equidad en la inteligencia artificial (IA) a través de la comprensión de los resultados del modelo. El artículo se compromete a explorar diversos métodos y marcos para la monitorización y explicabilidad de los modelos de ML.
91 IEEE	Fuzzy Rule-Based Explainer Systems for Deep Neural Networks: From Local Explainability to Global Understanding	Aghaeipoor, F., Sabokrou, M., & Fernández, A. (2023). Fuzzy Rule-Based Explainer Systems for Deep Neural Networks: From Local Explainability to Global Understanding. *IEEE Transactions on Fuzzy Systems*, 1–12. https://doi.org/10.1109/TFUZZ.2023	Este artículo propone sistemas de explicación basados en reglas difusas para redes neuronales profundas (FRBES). El algoritmo aprende un conjunto compacto pero preciso de reglas difusas basadas en la importancia de las características (es decir, valores de atribución) extraídas de las redes entrenadas. Estos sistemas pueden ser utilizados tanto para la explicabilidad local como global. Los resultados de la evaluación en diferentes aplicaciones

		.3243935	revelaron que los explicadores difusos mantenían la fidelidad y precisión de las redes neuronales profundas originales, a la vez que implican una menor complejidad y mejor comprensibilidad.
92 ACM	Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence	Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion, 99*, 101805. https://doi.org/10.1016/j.inffus.2023.10180	Este estudio ofrece una visión general de la investigación y tendencias actuales en XAI, incluyendo un estudio de caso. Se explica el trasfondo de XAI, sus definiciones comunes y se resumen las técnicas recientes propuestas para el aprendizaje supervisado. Se dividen las técnicas de XAI en cuatro ejes: explicabilidad de datos, explicabilidad de modelos, explicabilidad post-hoc y evaluación de explicaciones. Además, se presentan métricas de evaluación, paquetes de código abierto y conjuntos de datos, junto con direcciones futuras de investigación. El artículo destaca la importancia de la explicabilidad en términos de demandas legales, perspectivas de los usuarios y orientación de aplicaciones, proponiendo adaptar el contenido de las explicaciones a tipos de usuarios específicos. La evaluación de técnicas de XAI se realizó revisando 410 artículos críticos publicados entre enero de 2016 y octubre de 2022. El artículo está dirigido a investigadores de XAI interesados en hacer sus modelos de AI más confiables y a investigadores de otras disciplinas que buscan métodos efectivos de XAI.
93 ACM	Entropy-Based Logic Explanations of Neural Networks.	Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., & Melacci, S. (2022). Entropy-Based Logic Explanations of Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(6), 6046–6054. https://doi.org/10.1609/aaai.v36i6.20551	El método se basa en un criterio de entropía que identifica automáticamente los conceptos más relevantes. Se presentan cuatro estudios de caso para demostrar que: (i) este criterio basado en entropía facilita las explicaciones lógicas concisas en dominios críticos, desde datos clínicos hasta visión por computadora; (ii) el enfoque propuesto supera a los modelos de caja blanca de última generación en términos de precisión de clasificación y equipara el rendimiento de los modelos de caja negra.
94 ACM	Logic Explained Networks	Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., & Melacci, S. (2023). Logic Explained Networks. *Artificial Intelligence*, *314*, 103822. https://doi.org/10.1016/j.artint.2022.103822	En este artículo, se propone un enfoque general para la Inteligencia Artificial Explicable en el caso de arquitecturas neuronales, demostrando cómo una configuración estratégica de las redes conduce a una familia de modelos interpretables de aprendizaje profundo llamados Logic Explained Networks (LENs). Los LENs requieren que sus entradas sean predicados comprensibles para los humanos y proporcionan explicaciones en términos de fórmulas simples de Lógica de Primer Orden (FOL) que involucran dichos predicados. Los LENs son lo suficientemente generales como para cubrir una amplia gama de escenarios. Entre ellos, se consideran casos en los que los LENs se utilizan directamente como clasificadores especiales con capacidad explicativa, o actúan como redes adicionales que permiten explicar un clasificador de caja negra mediante fórmulas de FOL. Además de los problemas de aprendizaje

			supervisado, los LENs pueden aprender y proporcionar explicaciones en configuraciones de aprendizaje no supervisado. Resultados experimentales en diversos conjuntos de datos y tareas muestran que los LENs pueden producir clasificaciones superiores a modelos de caja blanca establecidos, como árboles de decisión y listas de reglas bayesianas, al mismo tiempo que ofrecen explicaciones más completas y significativas.
95 IEEE	Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks.	Nielsen, I. E., Dera, D., Rasool, G., Bouaynaya, N., & Ramachandran, R. P. (2022). Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks. *IEEE Signal Processing Magazine*, *39*(4), 73–84. https://doi.org/10.1109/MSP.2022.3142719	El paper se enfoca en los métodos de interpretación basados en gradientes, atribuyendo la responsabilidad de las decisiones a las características de entrada. Además, se discute la evaluación de estos métodos en términos de robustez, para obtener explicaciones significativas. Se analizan también las limitaciones y se proporcionan prácticas recomendadas para elegir un método de explicabilidad.
96 ACM	Deep Learning with Logical Constraints	Giunchiglia, E., Stoian, M. C., & Lukasiewicz, T. (2022). *Deep Learning with Logical Constraints* (arXiv:2205.00523). arXiv. http://arxiv.org/abs/2205.00523	El paper analiza cómo integrar conocimientos previos expresados en lógica de primer orden (FOL) en modelos de aprendizaje profundo para mejorar su rendimiento y explicabilidad.
97 IEEE	Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks.	Nielsen, I. E., Dera, D., Rasool, G., Bouaynaya, N., & Ramachandran, R. P. (2022). Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks. *IEEE Signal Processing Magazine*, *39*(4), 73–84. https://doi.org/10.1109/MSP.2022.3142719	Se presentan métodos de interpretabilidad, que utilizan el gradiente para asignar la responsabilidad de la decisión a las características de entrada.
98 ACM	Explaining the Black-box Smoothly - A Counterfactual Approach	Singla, S., Eslami, M., Pollack, B., Wallace, S., & Batmanghelich, K. (2023). Explaining the black-box smoothly—A counterfactual approach. *Medical Image Analysis, 84*, 102721. https://doi.org/10.1016/j.media.2022.102721	El artículo presenta el "BlackBox Counterfactual Explainer", un modelo diseñado para explicar las decisiones de clasificación de imágenes médicas. Los enfoques tradicionales, como los mapas de saliencia, no explican cómo las características de las imágenes en regiones anatómicas importantes son relevantes para la decisión de clasificación. Se realizó un experimento con residentes de radiología diagnóstica para comparar diferentes estilos de explicaciones (sin explicación, mapa de saliencia, explicación cycleGAN y nuestra explicación contrafactual) evaluando varios aspectos: comprensibilidad, justificación de la decisión del clasificador, calidad visual, preservación de la identidad y utilidad general de la explicación para los usuarios. Los resultados mostraron que la explicación contrafactual fue el único método que mejoró significativamente la comprensión de los usuarios sobre la decisión del

			clasificador en comparación con la línea base sin explicación.
99 Google Scholar	POST-HOC CONCEPT BOTTLENECK MODELS	Yuksekgonul, M., Wang, M., & Zou, J.Y. (2022). Post-hoc Concept Bottleneck Models. *ArXiv, abs/2205.15480*.	Los Modelos de Cuello de Botella Conceptual (CBMs) mejoran la interpretación, revelando que conceptos son importantes para una predicción sin sacrificar la precisión.
100 ACM	Supervised contrastive learning for interpretable long-form document matching	Jha, A., Rakesh, V., Chandrashekar, J., Samavedhi, A., & Reddy, C. K. (2023). Supervised contrastive learning for interpretable long-form document matching. *ACM Transactions on Knowledge Discovery from Data, 17*(2), Article 27. https://doi.org/10.1145/3542822	El artículo describe un modelo llamado CoLDE (Contrastive Long Document Encoder) diseñado para mejorar la correspondencia semántica entre documentos largos, como artículos científicos, documentos legales y patentes. Los modelos actuales se centran en documentos cortos y enfrentan dificultades con documentos largos debido a: Diferentes contextos para la misma palabra a lo largo del documento. Pequeñas secciones de texto similar entre documentos, pero diferencias en otras partes. La naturaleza general de una sola medida de similitud que no captura la heterogeneidad del contenido. CoLDE aborda estos problemas utilizando una estructura basada en transformadores con incrustaciones posicionales únicas y una capa de atención por fragmentos junto con un aprendizaje contrastivo supervisado. Captura la similitud en tres niveles: Puntuaciones de similitud a nivel general entre documentos. Puntuaciones de similitud entre diferentes secciones dentro y entre documentos. Puntuaciones de similitud entre diferentes fragmentos dentro del mismo documento y entre otros documentos. Estos puntajes detallados mejoran la interpretabilidad. CoLDE se evaluó en tres conjuntos de datos de documentos largos (publicaciones del ACL Anthology, artículos de Wikipedia y patentes del USPTO) y superó a los métodos actuales en la tarea de correspondencia de documentos, mostrando robustez ante cambios en la longitud del documento y perturbaciones del texto. El código del modelo está disponible públicamente.
101 ACM	A Question-centric Multi-experts Contrastive Learning Framework for Improving the Accuracy and Interpretability of Deep Sequential Knowledge Tracing Models	Zhang, H., Liu, Z., Shang, C., Li, D., & Jiang, Y. (2024). A Question-centric Multi-experts Contrastive Learning Framework for Improving the Accuracy and Interpretability of Deep Sequential Knowledge Tracing Models. *ACM Transactions on Knowledge Discovery from Data*, Article 3674840. https://doi.org/10.1145/3674840	El rastreo del conocimiento (KT) es importante para predecir el rendimiento futuro de los estudiantes mediante el análisis de sus procesos de aprendizaje históricos. Aunque las redes neuronales profundas (DNN) han mostrado potencial en resolver el problema de KT, enfrentan desafíos significativos: Modelar la información individual de las preguntas, ya que la adquisición de conocimiento por parte de los estudiantes puede variar considerablemente entre preguntas con los mismos componentes de conocimiento (KCs). Interpretar los resultados de predicción de los modelos de

			KT basados en aprendizaje profundo, lo cual es necesario para que los profesores comprendan y utilicen estos resultados en estrategias educativas. Para abordar estos desafíos, se propone un marco de aprendizaje contrastivo multi-expertos centrado en preguntas, llamado Q-MCKT. Este marco modela el estado de adquisición de conocimiento de los estudiantes a nivel de preguntas y conceptos, utilizando una técnica de mezcla de expertos para capturar un estado más robusto y preciso. Además, introduce una tarea de aprendizaje contrastivo centrada en preguntas para mejorar las representaciones de las preguntas con menos interacciones y utiliza una capa de predicción basada en la teoría de respuesta al ítem para generar resultados interpretables.
102 ACM	Knowledge graphs as tools for explainable machine learning: A survey	Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence, 302*, 103627. https://doi.org/10.1016/j.artint.2021.103627	Se propone el uso de grafos de conocimiento en el contexto del Aprendizaje Automático Explicable. La investigación en IA explicativa está creciendo al abordar las limitaciones de los modelos de aprendizaje automático actuales, que son precisos pero difíciles de interpretar. Se explora la integración de técnicas de Representación del Conocimiento en el Aprendizaje Automático para crear sistemas híbridos inteligentes. Se propone que los grafos de conocimiento, que proporcionan información de dominio en un formato legible por máquina, podrían mejorar las explicaciones ofrecidas por estos enfoques explicables. Utilizando una revisión sistemática de la literatura, se presenta un marco analítico para evaluar cómo los sistemas explicativos basados en conocimiento funcionan en diferentes dominios del aprendizaje automático. Se destacan las fortalezas de estos sistemas híbridos, como la mayor comprensión y precisión, así como sus limitaciones, y se concluye con una discusión sobre los desafíos abiertos para futuras investigaciones.
103 ACM	Interpretable local concept-based explanation with human feedback to predict all-cause mortality.	EL Shawi, R., & Al-Mallah, M. H. (2022). Interpretable local concept-based explanation with human feedback to predict all-cause mortality. *Journal of Artificial Intelligence Research, 75*. https://doi.org/10.1613/jair.1.14019	Marco de explicabilidad local basado en conceptos con retroalimentación humana (CLEF) para predecir la mortalidad por cualquier causa. El CLEF es un enfoque novedoso y agnóstico al modelo que utiliza conceptos etiquetados por clínicos en lugar de características crudas. Mapea las características de entrada a conceptos intuitivos de alto nivel y descompone la evidencia de la predicción en estos conceptos. Además, genera explicaciones contrafactuales que sugieren los cambios mínimos en la explicación basada en conceptos que llevarían a una predicción diferente. El estudio muestra que la retroalimentación directa de los usuarios es más efectiva que otras técnicas para alinear los conceptos aprendidos con las definiciones de conceptos de la realidad.

Bibliografía

Referencias Bibliográficas.

Abbott, A., & Callaway, E. (2014). Nobel prize for decoding brain's sense of place. Nature News, 514(7521), 153.

Aghaeipoor, F., Sabokrou, M., & Fernández, A. (2023). Fuzzy Rule-Based Explainer Systems for Deep Neural Networks: From Local Explainability to Global Understanding. IEEE Transactions on Fuzzy Systems, 1–12. https://doi.org/10.1109/TFUZZ.2023.3243935

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB (Vol. 1215, pp. 487–499).

Agughasi, V. I., & Murali, S. (Autores). (Fecha de publicación no proporcionada). xAI: An Explainable AI Model for the Diagnosis of COPD from CXR Images. En el Departamento de Ciencias de la Computación e Ingeniería, Maharaja Institute of Technology Mysore, Mysuru, India. [Correo electrónico de contacto: victor.agughasi@gmail.com, murali@mitmysore.in] [Enlace ORCID: https://orcid.org/0000-0002-1175-3089

Alape Suryanarayana, S., Sarne, D., & Kraus, S. (2022). Explainability in Mechanism Design: Recent Advances and the Road Ahead. arXiv preprint arXiv:2206.03031. [https://doi.org/10.48550/arXiv.2206.03031

Alape Suryanarayana, S., Sarne, D., & Kraus, S. (2022). Explainability in Mechanism Design: Recent Advances and the Road Ahead. ArXiv Preprint arXiv:2206.03031. Recuperado de https://doi.org/10.48550/arXiv.2206.03031

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Allen, G. I., Gan, L., & Zheng, L. (2024). Interpretable machine learning for discovery: Statistical challenges and opportunities. Annual Review of Statistics and Its Application, 11, 97-121. https://doi.org/10.1146/annurev-statistics-040120-030919

Almoussawi, Z. A., Kurdi, W. H. M., Khaleel, B. M., AL-Attabi, K., Sabah, H. A., & Alazzai, W. K. (2023). Planet Optimization with Machine Learning Enabled Power Usage Forecasting Modeling in Smart Grid Environment. En 6th International Conference on Engineering Technology and its Applications (IICETA) (pp. 726-732). Al-Najaf, Iraq. doi: 10.1109/IICETA57613.2023.10351470

Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2021). Explaining anomalies detected by autoencoders using shapley additive explanations. Expert Systems with Applications, 186, 115736. https://doi.org/10.1016/j.eswa.2021.115736

Arnold, V., Collier, P. A., Leech, S. A., & Sutton, S. G. (2004). Impact of intelligent decision aids on expert and novice decision-makers' judgments. Accounting & Finance, 44(1), 1–26.

Asakawa, C. (2023). Interaction Techniques with a Navigation Robot for the Visually Impaired. In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23) (p. 1). https://doi.org/10.1145/3568162.3576952

Badreddine, S., d'Avila Garcez, A., Serafini, L., & Spranger, M. (2022). Logic tensor networks. Artificial Intelligence, 303.

Baker, S., & Xiang, W. (2023). Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence. Journal Name. Retrieved from https://www.researchgate.net/profile/Stephanie-Baker-12/publication/376412417_Explainable_AI_is_Responsible_AI_How_Explainability_Creates_Trustworthy_and_Socially_Responsible_Artificial_Intelligence/links/6577d1cdfc4b416622b8b444/Explainable-AI-is-Responsible-AI-How-Explainability-Creates-Trustworthy-and-Socially-Responsible-Artificial-Intelligence.pdf

Balestra, C., Li, B., & Müller, E. (2023). slidSHAPs – sliding Shapley Values for correlation-based change detection in time series. In Proceedings of the 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-10). Thessaloniki, Greece. doi: 10.1109/DSAA60987.2023.10302636 .

Balla, Y., Tirunagari, S., & Windridge, D. (2023). Pediatrics in Artificial Intelligence Era: A Systematic Review on Challenges, Opportunities, and Explainability. Indian Pediatrics, 60, 561–569. https://doi.org/10.1007/s13312-023-2936-8

Balla, Y., Tirunagari, S., & Windridge, D. (2023, May 14). Machine Learning in Pediatrics: Evaluating Challenges, Opportunities, and Explainability. PMID: 37179470. Advance online publication. https://doi.org/S097475591600533

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel vision computation. Nature.

Barnard, P., Marchetti, N., & DaSilva, L. A. (2022). Robust network intrusion detection through explainable artificial intelligence (XAI). IEEE Networking Letters, 4(3), 167–171. https://doi.org/10.1109/LNET.2022.3197355

Barragán-Montero, A., Bibal, A., Dastarac, M., Draguet, C., Valdes, G., Nguyen, D., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Souris, K., Sterpin, E., & Lee, J. (2022). Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. Physics in Medicine & Biology, 67. https://doi.org/10.1088/1361-6560/ac678a

Başağaoğlu, H., Chakraborty, D., Do Lago, C., Gutierrez, L., Şahinli, M. A., Giacomoni, M., Furl, C., Mirchi, A., Moriasi, D., & Şengör, S. S. (2022). A review on interpretable and explainable artificial intelligence in hydroclimatic applications. Water, 14(8), 1230. https://doi.org/10.3390/w14081230

Basu, S., Majumdar, B., Mukherjee, K., Munjal, S., & Palaksha, C. (2023). Artificial Intelligence–HRM Interactions and Outcomes: A Systematic Review and Causal Configurational Explanation. Human Resource Management Review, 33(1), 100893. https://doi.org/10.1016/j.hrmr.2022.100893

Baur, L., Ditschuneit, K., Schambach, M., Kaymakci, C., Wollmann, T., & Sauer, A. (2024). Explainability and Interpretability in Electric Load Forecasting Using Machine Learning Techniques – A Review. Energy AI, 1, 100358. https://doi.org/10.1016/j.egyai.2024.100358

Begum, M., Alam, M., Islam, M. R., & Hossain, M. A. (2024). LCNN: Lightweight CNN Architecture for Software Defect Feature Identification Using Explainable AI. IEEE Access, 12, 55744-55756. doi: 10.1109/ACCESS.2024.3388489 .

Bhatt, H. S., Ramakrishnan, S., Raja, S., & Jawahar, C. V. (2024). Unlocking the potential of unstructured data in business documents through document intelligence. En Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (pp. 505–509). https://doi.org/10.1145/3632410.3633293

Biran, O., & McKeown, K. R. (2017). Human-centric justification of machine learning predictions. In IJCAI.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1989). Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM (JACM), 36(4), 929–965.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176.

Cañete-Sifuentes, L., Robles, V., Menasalvas, E., & Monroy, R. (2023). Comparing Automated Machine Learning Against an Off-the-Shelf Pattern-Based Classifier in a Class Imbalance Problem: Predicting University Dropout. IEEE Access, 11, 139147-139156. doi: 10.1109/ACCESS.2023.3336596 .

Chakrabarti, A., Patra, A., & Noble, J. A. (2020). Contrastive fairness in machine learning. IEEE Letters of the Computer Society, 3(2), 38-41.

Champendal, M., Müller, H., Prior, J. O., & dos Reis, C. S. (2023). A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. European Journal of Radiology, 111159. https://doi.org/10.1016/j.ejrad.2023.111159

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., & Rudin, C. (2018). This Looks Like That: Deep Learning for Interpretable Image Recognition. CoRR. Retrieved from https://arxiv.org/abs/1806.10574

Chen, Z., Bei, Y., & Rudin, C. (2020). Concept whitening for interpretable image recognition. Nature Machine Intelligence, 2(12), 772-782

Choi, S., Kim, T., Oh, K., Lee, E., Kim, D., & Park, J. (2020). Diagnóstico del infarto agudo de miocardio basado en una red neuronal convolucional: estudio retrospectivo observacional. Revista Internacional de Investigación Ambiental y Salud Pública, 17(16), 5693.

Ciobanu-Caraus, O., Aicher, A., Kernbach, J.M. et al. A critical moment in machine learning in medicine: on reproducible and interpretable learning. Acta Neurochir 166, 14 (2024). https://doi.org/10.1007/s00701-024-05892-8

Ciravegna, G., Giannini, F., Gori, M., Maggini, M., & Melacci, S. (2020). Human-driven FOL explanations of deep learning. In Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20} (pp. 2234-2240). International Joint Conferences on Artificial Intelligence Organization .

Conard, A. M., DenAdel, A., & Crawford, L. (2023). A spectrum of explainable and interpretable machine learning approaches for genomic studies. Wiley Interdisciplinary Reviews: Computational Statistics, 15, e1617. https://doi.org/10.1002/wics.1617

Cui, S., Traverso, A., Niraula, D., Zou, J., Luo, Y., Owen, D., El Naqa, I., Wei, L. (2023). Interpretable artificial intelligence in radiology and radiation oncology. British Journal of Radiology, 96(1150), 20230142. https://doi.org/10.1259/bjr.20230142

Dabiri, S., Beheshti, F., Ghassemi, N., Mirniaharikandehei, S., & Rafiee, V. (2019). Evaluación de la segmentación del miocardio utilizando Grad-CAM en la resonancia magnética cardíaca de perfusión. IEEE Transactions on Medical Imaging, 38(9), 2201-2212.

Dai, W.-Z., Xu, Q., Yu, Y., & Zhou, Z.-H. (2019). Bridging machine learning and logical reasoning by abductive learning. En Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 2019.

Das, D., Kim, B., & Chernova, S. (2023). Subgoal-based explanations for unreliable intelligent decision support systems. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23) (pp. 240–250). https://doi.org/10.1145/3581641.3584055

Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G., Durand, F., & Freeman, W. T. (2014). The visual microphone: Passive recovery of sound from video. ACM Transactions on Graphics (Proc. SIGGRAPH), 33(4), 79:1–79:10.

Demirović, E., Lukina, A., Hebrard, E., Chan, J., Bailey, J., Leckie, C., Ramamohanarao, K., & Stuckey, P. J. (2022). MurTree: Optimal decision trees via Dynamic programming and search. The Journal of Machine Learning Research, 23(1), 1169–1215.https://dl.acm.org/doi/10.5555/3586589.3586615

Dennis, D. K., Li, T., & Smith, V. (2021). Heterogeneity for the win: One-shot federated clustering. In M. Meila & T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 2611-2620). Retrieved from https://proceedings.mlr.press/v139/dennis21a.html

Denton, E., Chintala, S., Szlam, A., & Fergus, R. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. NIPS.

Diligenti, M., Gori, M., Maggini, M., & Rigutini, L. (2012). Bridging logic and kernel machines. Machine Learning, 86.

Diligenti, M., Roychowdhury, S., & Gori, M. (2017). Integrating prior knowledge into deep learning. En Proceedings of the International Conference on Machine Learning and Applications (ICMLA), 2017.

Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. Information Sciences. Advance online publication. https://doi.org/10.1016/j.ins.2022.10.013

Donadello, I., Serafini, L., & D'Avila Garcez, A. (2017). Logic tensor networks for semantic image interpretation. En Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2017.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dubey, S. A., & Pandit, A. A. (2022). A comprehensive review and application of interpretable deep learning model for ADR prediction. International Journal of Advanced Computer Science and Applications (IJACSA), 13(9). http://dx.doi.org/10.14569/IJACSA.2022.0130924

EL Shawi, R., & Al-Mallah, M. H. (2022). Interpretable local concept-based explanation with human feedback to predict all-cause mortality. Journal of Artificial Intelligence Research, 75. https://doi.org/10.1613/jair.1.14019

El Zini, J., & Awad, M. (2023). On the explainability of natural language processing deep models. ACM Computing Surveys, 55(5), Article 103, 1–31. https://doi.org/10.1145/3529755

Etienne, H. (2022). Computational philosophy. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22) (p. 899). https://doi.org/10.1145/3514094.3539562

Farah, L., Murris, J. M., Borget, I., Guilloux, A., Martelli, N. M., & Katsahian, S. I. M. (2023). Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence–Based Health Technologies: What Healthcare Stakeholders Need to Know. Mayo Clinic Proceedings: Digital Health, 1(2), 120-138. https://doi.org/10.1016/j.mcpdig.2023.02.004

Farhadloo, M., Molnar, C., Luo, G., Li, Y., Shekhar, S., Maus, R. L., Markovic, S., Leontovich, A., & Moore, R. (2022). SAMCNet: Towards a Spatially Explainable AI Approach for Classifying MxIF Oncology Data. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), (pp. 2860–2870). https://doi.org/10.1145/3534678.3539168

Feng, J., Shaib, C., & Rudzicz, F. (2020). Explainable clinical decision support from text. In EMNLP (pp. 1478-1489).

Fischer, G., Mastaglio, T., Reeves, B., & Rieman, J. (1990). Minimalist explanations in knowledge-based systems. In Twenty-Third Annual Hawaii International Conference on System Sciences, volume 3 (pp. 309–317 vol.3).

Flora, M., Potvin, C., McGovern, A., & Handler, S. (2022, November 16). Comparing Explanation Methods for Traditional Machine Learning Models Part 1: An Overview of Current Methods and Quantifying Their Disagreement. doi:10.48550/arXiv.2211.08943

Flores-Araiza, D., et al. (2023). Deep Prototypical-Parts Ease Morphological Kidney Stone Identification and are Competitively Robust to Photometric Perturbations. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 295-304). Vancouver, BC, Canada. doi: 10.1109/CVPRW59228.2023.00035 .

Flores-Araiza, D., Lopez-Tiro, F., Villalvazo-Avila, E., El-Beze, J., Hubert, J., Ochoa-Ruiz, G., & Daul, C. (2022). Interpretable deep learning classifier by detection of prototypical parts on kidney stones images. arXiv preprint arXiv:2206.00252.

Francisco Gutiérrez, Xavier Ochoa, Karsten Seipp, Tom Broos, and Katrien Verbert. 2019. Benefits and Trade-Offs of Different Model Representations in Decision Support Systems for Non-expert Users. In INTERACT, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, 576–597

Frasca, M., La Torre, D., Pravettoni, G., et al. (2024). Explainable and interpretable artificial intelligence in medicine: A systematic bibliometric review. Discovery of Artificial Intelligence, 4(15), 1-14. https://doi.org/10.1007/s44163-024-00114-7

Freitas, A. A. (2014). Comprehensible classification models: A position paper. ACM SIGKDD Explorations Newsletter, 15(1), 1-10.

Gao, R., Xie, J., Zhu, S.-C., & Wu, Y. N. (2018). Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. In International Conference on Learning Representations (ICLR).

Giacobbe, D. R., Marelli, C., Guastavino, S., Mora, S., Rosso, N., Signori, A., Campi, C., Giacomini, M., & Bassetti, M. (2024). Explainable and interpretable machine learning for antimicrobial stewardship: Opportunities and challenges. Clinical Therapeutics. Advance online publication. https://doi.org/10.1016/j.clinthera.2024.02.010

Giunchiglia, E., Stoian, M. C., & Lukasiewicz, T. (2022). Deep Learning with Logical Constraints (arXiv:2205.00523). arXiv. http://arxiv.org/abs/2205.00523

Gkatzia, D., Lemon, O., & Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. arXiv preprint arXiv:1606.03254 .

Goodfellow, I., Bengio, Y., & Courville, A. (2017). Deep Learning. MIT Press.

Goodwin, N. L., Nilsson, S. R. O., Choong, J. J., & Golden, S. A. (2022). Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. Neurobiology of Behavior 2022, Edited by Tiago Branco and Mala Murthy. University of Washington.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), Article 93. https://doi.org/10.1145/3236009

Gunning, D. (2017). Explainable Artificial Intelligence (XAI). Defence Advanced Research Projects Agency (DARPA). Recuperado de https://www.darpa.mil/attachments/DARPA_XAI_Program_Summary.pdf

Guo, H., Jia, F., Chen, J., Squicciarini, A., & Yadav, A. (2023). RoCourseNet: Robust Training of a Prediction Aware Recourse Model. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), 619–628. https://doi.org/10.1145/3583780.3615040

Guo, H., Nguyen, T., & Yadav, A. (2021). CounterNet: End-to-end training of counterfactual aware predictions. In ICML 2021 Workshop on Algorithmic Recourse.

Guo, Z., Shen, Y., Wan, S., Shang, W. L., & Yu, K. (2022). Hybrid Intelligence-Driven Medical Image Recognition for Remote Patient Diagnosis in Internet of Medical Things. IEEE Journal of Biomedical and Health Informatics, 26(12), 5817-5828. doi:10.1109/JBHI.2021.3139541 .

Gutiérrez, F., Ochoa, X., Seipp, K., Broos, T., & Verbert, K. (2019). Benefits and trade-offs of different model representations in decision support systems for non-expert users. In D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, & P. Zaphiris (Eds.), INTERACT (pp. 576-597). Springer International Publishing.

Haffar, R., Sánchez, D., & Domingo-Ferrer, J. (2022). Explaining predictions and attacks in federated learning via random forests. Applied Intelligence, 53(1), 169–185. https://doi.org/10.1007/s10489-022-03435-1

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: Some applications. Journal of the American Statistical Association, 82(398), 371-386 .

Henriques, J., Rocha, T., de Carvalho, P., Silva, C., Paredes, S. (2024). Interpretability and Explainability of Machine Learning Models: Achievements and Challenges. En: Pino, E., Magjarević, R., de Carvalho, P. (eds) Conferencia Internacional sobre Informática Biomédica y de Salud 2022. ICBHI 2022. Actas de IFMBE, vol. 108. Springer, Cham. https://doi.org/10.1007/978-3-031-59216-4_9

Hu, Z., Ma, X., Liu, Z., Hovy, E., & Xing, E. (2016a). Harnessing deep neural networks with logic rules. En Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2016.

Hu, Z., Yang, Z., Salakhutdinov, R., & Xing, E. (2016b). Deep neural networks with massive learned knowledge. En Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016.

Huong, T. T., Bac, T. P., Ha, K. N., Hoang, N. V., Hoang, N. X., Hung, N. T., & Tran, K. P. (2022). Federated learning-based explainable anomaly detection for industrial control systems. IEEE Access, 10, 53,854–53,872.

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. Applied Sciences, 12(3), 1353. https://doi.org/10.3390/app12031353

Jacobs, M. (2024). Alternative interpretable machine learning models applied to corporate probability of default: A literature review and high points of a benchmarking analysis. SSRN. https://doi.org/10.2139/ssrn.4583014

Janah, N. Z., Permanasari, A. E., & Setiawan, N. A. (2023). Phase lag index of visual-memory processing EEG for computer-aided AUD diagnosis. In Proceedings of the 2023 9th International Conference on Computer Technology Applications (ICCTA '23) (pp. 143–150). https://doi.org/10.1145/3605423.3605452

Jha, A., Rakesh, V., Chandrashekar, J., Samavedhi, A., & Reddy, C. K. (2023). Supervised contrastive learning for interpretable long-form document matching. ACM Transactions on Knowledge Discovery from Data, 17(2), Article 27. https://doi.org/10.1145/3542822

Jones, R. W., Mateer, J. E., & Harrison, M. J. (2019). Malfunction transparency in clinical decision support systems: A classification approach. In ICIEA. IEEE, 1354–1359.

Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. College of Nursing, Seoul National University, Seoul, Republic of Korea. Department of Computer

Science and Engineering, University of Seoul, Seoul, Republic of Korea. Department of Artificial Intelligence, University of Seoul, Seoul, Republic of Korea. Emergency Nursing Department, Seoul National University Hospital, Seoul, Republic of Korea. Research Institute of Nursing Science, College of Nursing, Seoul National University, Seoul, Republic of Korea. Center for Human-Caring Nurse Leaders for the Future by Brain Korea 21 (BK 21) Four Project, College of Nursing, Seoul National University, Seoul, Republic of Korea. Received 6 November 2022, Revised 26 March 2023, Accepted 5 May 2023, Available online 8 May 2023, Version of Record 16 May 2023 .

Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon. 9. e16110. https://doi.org/10.1016/j.heliyon.2023.e16110 .

Kadir, M. A., Mosavi, A., & Sonntag, D. (2023, May 5). Assessing XAI: Unveiling Evaluation Metrics for Local Explanation, Taxonomies, Key Concepts, and Practical Applications. German Research Center for Artificial Intelligence & John von Neumann Faculty of Informatics, Obuda University .

Kang, B., Kweon, J., Rangaswamy, M., & Monga, V. (2023). Deep Learning for Radar Waveform Design: Retrospectives and the Road Ahead. In IEEE International Radar Conference (RADAR) (pp. 1-6). Sydney, Australia. doi: 10.1109/RADAR54928.2023.10371126 .

Karim, M. R., et al. (2023). Interpreting black-box machine learning models for high dimensional datasets. In Proceedings of the 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-10). IEEE. https://doi.org/10.1109/DSAA60987.2023.10302562

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. arXiv preprint arXiv:2010.04050.

Karval, R., & Singh, K. N. (2023). Catching Silent Failures: A Machine Learning Model Monitoring and Explainability Survey. In 2023 OITS International Conference on Information Technology (OCIT) (pp. 526-532). Raipur, India. doi: 10.1109/OCIT59427.2023.10431343 .

Kass, R., & Finin, T. (1988). The Need for User Models in Generating Expert System Explanations. International Journal of Expert Systems, 1(4).

Kazhdan, D., Dimanov, B., Jamnik, M., Lio, P., & Weller, A. (2020). Now You See Me (CME): Concept-based Model Extraction. arXiv preprint arXiv:2010.13233

Keany, J., McDowell, K., & Quesada, M. (2019). Diferenciación entre espirales y otras manchas blancas en imágenes de resonancia magnética por inteligencia artificial: un estudio de prueba de concepto. European Heart Journal - Cardiovascular Imaging, 20(5), 514-520.

Kim, B., Gilmer, J., Wattenberg, M., & Viegas, F. (2018). Tcav: Relative concept importance testing with linear concept activation vectors.

Kitamura, K., Irvan, M., & Shigetomi Yamaguchi, R. (2023). XAI for Medicine by ChatGPT Code interpreter. In Proceedings of the 2023 5th International Conference on Big-data Service and Intelligent Computation (pp. 28–34). https://doi.org/10.1145/3633624.3633629

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In International Conference on Machine Learning (pp. 5338-5348). PMLR.

Kononenko, I., et al. (2010). An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11(Jan), 1–18.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. arXiv preprint arXiv:1703.06856.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1675–1684). ACM.

LeCun, Y. (1989). Generalization and network design strategies. Technical Report CRG-TR-89-4, University of Toronto.

Li, Y., Farhadloo, M., Krishnan, S., Frankel, T. L., Shekhar, S., & Rao, A. (2021). SRNet: A spatial-relationship aware point-set classification method for multiplexed pathology images. In Proceedings of DeepSpatial'21, Vol. 10.

Lin, C., Lin, C. M., Lin, B., & Yang, M. C. (2009). A decision support system for improving doctors' prescribing behavior. Expert Systems with Applications, 36(4), 7975–7984. https://doi.org/10.1016/j.eswa.2008.10.045

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

Liu, H., Zhong, C., Alnusair, A., & Islam, S. R. (2021). Faixid: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques. Journal of Network and Systems Management, 29(4), 40. https://doi.org/10.1007/s10922-021-09606-8

Liu, K., Faraji Niri, M., Apachitei, G., Lain, M., Greenwood, D., & Marco, J. (2022). Interpretable machine learning for battery capacities prediction and coating parameters analysis. Control Engineering Practice, 123, 105202. https://doi.org/10.1016/j.conengprac.2022.105202

Liyanage, K. S. K., Tian, Z., Divakaran, D. M., Chan, M. C., & Gurusamy, M. (2022). Apex: Characterizing attack behaviors from network anomalies. In 2022 IEEE International Performance, Computing, and Communications Conference (IPCCC) (pp. 207–216). https://doi.org/10.1109/IPCCC59043.2022.9812912

Lombrozo, T. (2012). Explanation and abductive inference.

Lopez, F., Varelo, A., Hinojosa, O., Mendez, M., Trinh, D.-H., ElBeze, Y., Hubert, J., Estrade, V., Gonzalez, M., Ochoa, G., & et al. (2021). Assessing deep learning methods for the identification of kidney stones in endoscopic images. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 2778–2781). IEEE.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. En Advances in Neural Information Processing Systems, 30, 4765-4774

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874 .

Machado, J. P., Lam, X. T., & Chen, J.-W. (2018). Use of a clinical decision support tool for the management of traumatic dental injuries in the primary dentition by novice and expert clinicians. Dental Traumatology, 34(2), 120–128. https://doi.org/10.1111/edt.12385

Madapatha, S., & Fernando, P. (2024). A Systematic Literature Review of XAI-based Approaches on Brain Disease Detection using Brain MRI Images. In 4th International Conference on Advanced Research in Computing (ICARC) (pp. 19-24). Belihuloya, Sri Lanka. doi: 10.1109/ICARC61713.2024.10499752 .

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., & Lao, N. (2020). Multi-scale representation learning for spatial feature distributions using grid cells. In ICLR (2020). Retrieved from https://openreview.net/forum?id=BylRTeHFvr

Maley, C. C., Koelble, K., Natrajan, R., Aktipis, A., & Yuan, Y. (2015). An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer. Breast Cancer Research, 17(1), 1-13.

Marčinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. First published: 28 February 2023. Edited by: Mehmed Kantardzic, Associate Editor and Witold Pedrycz, Editor-in-Chief. DOI: https://doi.org/10.1002/widm.1493

Marcus, E., & Teuwen, J. (2024). Artificial intelligence and explanation: How, why, and when to explain black boxes. European Journal of Radiology, 173, 111393. https://doi.org/10.1016/j.ejrad.2024.111393

Marra, G., Giannini, F., Diligenti, M., & Gori, M. (2019). LYRICS: A general interface layer to integrate logic inference and deep learning. En Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2019.

Martin, W., Sheynkman, G., Lightstone, F. C., Nussinov, R., & Cheng, F. (2022). Interpretable artificial intelligence and exascale molecular dynamics simulations to reveal kinetics: Applications to Alzheimer's disease. Current Opinion in Structural Biology, 72, 103-113. https://doi.org/10.1016/j.sbi.2021.09.001

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. Journal of Mathematical Sociology, 4(1), 103-120 .

Mehdiyev, N., Majlatow, M., & Fettke, P. (2023). Interpretable and Explainable Machine Learning Methods for Predictive Process Monitoring: A Systematic Literature Review. arXiv preprint arXiv:2312.17584. https://doi.org/10.48550/arXiv.2312.17584

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

Minh, D., Wang, H. X., Li, Y. F., et al. (2022). Explainable artificial intelligence: A comprehensive review. Artificial Intelligence Review, 55, 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Mueller, S. T., Veinott, E. S., Hoffman, R. R., et al. (2021). Principles of explanation in human-AI systems. CoRR arXiv:2102.04972.

Mutlu, E. Ç., Yousefi, N., & Garibay, O. O. (2022). Contrastive counterfactual fairness in algorithmic decision-making. En Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22) (pp. 499–507). https://doi.org/10.1145/3514094.3534143

Nauta, M., Jutte, A., Provoost, J. C., & Seifert, C. (2020). This looks like that, because ... explaining prototypes for interpretable image recognition. CoRR, abs/2011.02863. Retrieved from https://arxiv.org/abs/2011.02863

Neshenko, N., Bou-Harb, E., Crichigno, J., Kaddoum, G., & Ghani, N. (2019). Demystifying IoT security: An exhaustive survey on IoT vulnerabilities and a first empirical look on internet-scale IoT exploitations. IEEE Communications Surveys Tutorials, 21(3), 2702–2733.

Nguyen, Q. P., Lim, K. W., Divakaran, D. M., Low, K. H., & Chan, M. C. (2019). Gee: A gradient-based explainable variational autoencoder for network anomaly detection. In 2019 IEEE Conference on Communications and Network Security (CNS) (pp. 91–99). IEEE. https://doi.org/10.1109/CNS48642.2019.8933011

Nguyen, T.-D. H., Bui, N., Nguyen, D., Yue, M.-C., & Nguyen, V. A. (2022). Robust Bayesian Recourse. In The 38th Conference on Uncertainty in Artificial Intelligence.

Nielsen, I. E., Dera, D., Rasool, G., Bouaynaya, N., & Ramachandran, R. P. (2022). Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks. IEEE Signal Processing Magazine, 39(4), 73–84. https://doi.org/10.1109/MSP.2022.3142719

Olsson, O. (2022). A Taxonomy of Non-Fungible Tokens: Overview, Evaluation and Explanation (Ph.D. Dissertation). Uppsala Universitet, Sweden. Order Number: AAI29353266 .

Ozalp, R., Ucar, A., & Guzelis, C. (2024). Advancements in Deep Reinforcement Learning and Inverse Reinforcement Learning for Robotic Manipulation: Toward Trustworthy, Interpretable, and Explainable Artificial Intelligence. IEEE Access, 12, 51840-51858. DOI: 10.1109/ACCESS.2024.3385426 .

Papamichail, K. N., & French, S. (2000). Decision support in nuclear emergencies. Journal of Hazardous Materials, 71(1-3), 321–342.

Pappa, G. L., Baines, A. J., & Freitas, A. A. (2005). Predicting post-synaptic activity in proteins with data mining. Bioinformatics, 21(suppl 2), ii19–ii25.

Park, M., & Lee, K. (2022). Exploiting negative preference in content-based music recommendation with contrastive learning. In Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22) (pp. 229–236). https://doi.org/10.1145/3523227.3546768

Park, R. G. (2022). A regional explanation for Laxfordian tectonic evolution and its implications for the Lewisian terrane model. Scottish Journal of Geology. https://doi.org/10.1144/sjg2021-020

Pashamokhtari, A., Batista, G., & Gharakheili, H. H. (Year). Efficient IoT Traffic Inference: From Multi-view Classification to Progressive Monitoring. ACM Transactions on Internet of Things, 5(1), 1–30. https://doi.org/10.1145/3625306

Peters, M., Neumann, M., Zettlemoyer, L., & Yih, W.-t. (2018). Dissecting contextual word embeddings: Architecture and representation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1499–1509.

Ponnusamy, U., D. D. B. S., & Sampathila, N. (2023). Approaching explainable artificial intelligence methods in the diagnosis of iron deficiency anemia using blood parameters. In Proceedings of the 2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS) (pp. 201-206). IEEE. https://doi.org/10.1109/ICRAIS59684.2023.10367126

Qiu, S., Xu, H., Deng, J., Jiang, S., & Lu, L. (2019). Transfer convolutional neural network for cross-project defect prediction. Applied Sciences, 9(13), 2660. https://doi.org/10.3390/app9132660

Quinlan, J. R. (1987a). Generating production rules from decision trees. In IJCAI (Vol. 87, pp. 304–307).

Quinlan, J. R. (1987b). Simplifying decision trees. International Journal of Man-Machine Studies, 27(3), 221–234.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Elsevier.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Ran, S.-J., & Su, G. (2023, November 17). Tensor Networks for Interpretable and Efficient Quantum-Inspired Machine Learning. Intelligent Computing, 2, Article ID: 0061. https://doi.org/10.34133/icomputing.0061

Rao, D., & Mane, S. (2021). Zero-shot learning approach to adaptive cybersecurity using explainable AI. CoRR, abs/2106.14647. Retrieved from https://arxiv.org/abs/2106.14647

Ren, Z., Qin, X., & Wang, B. (2023). Unraveling the impact of explainability of artificial intelligence-generated content (AIGC) on design style transfer effects. En Proceedings of the 2023 9th International Conference on Communication and Information Processing (ICCIP '23) (pp. 171–185). https://doi.org/10.1145/3638884.3638910

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM.

Riva, G., Sajno, E., De Gaspari, S., Pupillo, C., & Wiederhold, B. K. (Year). Navigating the Ethical Crossroads: Bridging the gap between Predictive Power and Explanation in the use of Artificial Intelligence in Medicine. Annual Review of CyberTherapy and Telemedicine, 21(N/A), 3-7. https://hdl.handle.net/10807/272879

Rivest, R. L. (1987). Learning decision lists. Machine Learning, 2(3), 229-246.

Rumelhart, D., Hinton, G., & Williams, R. (1986a). Learning representations by back-propagating errors. Nature, 323,533–536 .

Sáez-de-Cámara, X., Flores, J. L., Arellano, C., Urbieta, A., & Zurutuza, U. (2023). Federated Explainability for Network Anomaly Characterization. In Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '23) (pp. 346–365). https://doi.org/10.1145/3607199.3607234

Salih, A., Boscolo Galazzo, I., Gkontra, P., Lee, A. M., Lekadir, K., Raisi-Estabragh, Z., & Petersen, S. E. (2023). Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models. Circulation: Cardiovascular Imaging, 16. https://doi.org/10.1161/CIRCIMAGING.122.014519

Sasaki, H., & Takenouchi, T. (2022). Representation Learning for Maximization of MI, Nonlinear ICA and Nonlinear Subspaces with Robust Density Ratio Estimation. Journal of Machine Learning Research, 23(1), 1-55. Submitted 12/20; Revised 6/22; Published 8/22. https://dl.acm.org/doi/10.5555/3586589.3586820

Sato, M., & Tsukimoto, H. (2001). Rule extraction from neural networks via decision tree induction. In IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Vol. 3, pp. 1870-1875). IEEE.

Sazzed, S. (2022). Stylometric and Semantic Analysis of Demographically Diverse Non-native English Review Data. In 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 470-476). Istanbul, Turkey. https://doi.org/10.1109/ASONAM55673.2022.10068612

Schreyer, M., Sattarov, T., & Borth, D. (2021). Multi-view contrastive self-supervised learning of accounting data representations for downstream audit tasks. In Proceedings of the Second ACM International Conference on AI in Finance (ICAIF '21) (Article No. 8, pp. 1–8). https://doi.org/10.1145/3490354.3494373

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery. Advance online publication. https://doi.org/10.1007/s10618-022-00867-8

Schwalbe, G., Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min Knowl Disc (2023). https://doi.org/10.1007/s10618-022-00867-8

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization. arXiv preprint arXiv:1610.02391 .

Serafini, L., & d'Avila Garcez, A. (2016). Logic tensor networks: Deep learning and logical reasoning from data and knowledge. En Proceedings of the Workshop on Neural-Symbolic Learning and Reasoning (NeSy), colocada dentro de la International Joint Conference on Neural Networks (IJCNN), 2016.

Shahani, S., Abraham, J., & Venkateswaran. (Year). Techniques for Privacy-Preserving Data Aggregation in an Untrusted Distributed Environment. In Proceedings of the 6th Joint International Conference on Data Science & Management of Data (CODS-COMAD '23) (pp. 286–287). https://doi.org/10.1145/3570991.3571020

Shakeel Sheikh, T., Shim, J., & Cho, M. (2023). Clustering-based cancer diagnosis model for whole slide image. In Proceedings of the 2023 8th International Conference on Biomedical Imaging, Signal Processing (ICBSP '23) (pp. 1–8). https://doi.org/10.1145/3634875.3634876

Shan, S., Baskaran, V. A., Yi, H., Ranek, J., Stanley, N., & Oliva, J. B. (2022). Transparent single-cell set classification with kernel mean embeddings. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22) (pp. 1–10). https://doi.org/10.1145/3535508.3545538

Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. En International Symposium on Spatial and Temporal Databases (pp. 236-256). Springer.

Sheu, R.-K., & Pardeshi, M. S. (2022). A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. Sensors, 22(20), 8068. https://doi.org/10.3390/s22208068

Simon, H. A. (1975). The functional equivalence of problem solving skills. Cognitive Psychology, 7(2), 268-288.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Singla, S., Eslami, M., Pollack, B., Wallace, S., & Batmanghelich, K. (2023). Explaining the black-box smoothly—A counterfactual approach. Medical Image Analysis, 84, 102721. https://doi.org/10.1016/j.media.2022.102721

Sipos, L., Schäfer, U., Glinka, K., & Müller-Birn, C. (2023). Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank. In Mensch und Computer 2023 (MuC '23), September 03–06, 2023, Rapperswil, Switzerland (pp. 1–6). ACM, New York, NY, USA. https://doi.org/10.1145/3603555.3608551

Skopik, F., Landauer, M., & Wurzenberger, M. (2022). Blind spots of security monitoring in enterprise infrastructures: A survey. IEEE Security & Privacy, 20(6), 18–26.

Smith, J., & Johnson, A. (2023). The Pros and Cons of Using Machine Learning and Interpretable Machine Learning Methods In Psychiatry Detection Applications Specifically Depression Disorder: A Brief Review. ResearchGate. https://www.researchgate.net/publication/375331868/The_Pros_and_Cons_of_Using_Machine_Learning_and_Interpretable_Machine_Learning_Methods_In_Psychiatry_Detection_Applications_Specifically_Depression_Disorder_A_Brief_Review

Soon, R.-J., Sang, D. V., Chng, C.-B., & Chui, C.-K. (2023). Explainable AI for CPS-Based Manufacturing Workcell. In 2023 International Conference on System Science and

Engineering (ICSSE) (pp. 332-337). Ho Chi Minh, Vietnam. doi: 10.1109/ICSSE58758.2023.10227195

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) (pp. 2239–2250). https://doi.org/10.1145/3531146.3534639

Sudheera, K. L. K., Divakaran, D. M., Singh, R. P., & Gurusamy, M. (2021). Adept: Detection and identification of correlated attack stages in IoT networks. IEEE Internet of Things Journal, 8(8), 6591–6607. https://doi.org/10.1109/JIOT.2021.3117747

Tanuwidjaja, H. C., Takahashi, T., Lin, T.-N., Lee, B., & Ban, T. (2023). Hybrid explainable intrusion detection system: Global vs. local approach. In Proceedings of the 2023 Workshop on Recent Advances in Resilient and Trustworthy ML Systems in Autonomous Networks (ARTMAN '23) (pp. 37–42). https://doi.org/10.1145/3605772.3624004

Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A. P., & Bruder, C. (2023). Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work. En H. Degen & S. Ntoa (Eds.), Artificial Intelligence in HCI. HCII 2023. Lecture Notes in Computer Science (Vol. 14050). Springer, Cham. https://doi.org/10.1007/978-3-031-35891-3_22

Tian, Q., Gao, H., Guo, L., Li, Z., & Wang, Q. (2023, November). CSG compressive strength prediction based on LSTM and interpretable machine learning. Reviews on Advanced Materials Science, 62(1). DOI: 10.1515/rams-2023-0133

Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. Artificial Intelligence, 302, 103627. https://doi.org/10.1016/j.artint.2021.103627

Tong, H., Liu, B., & Wang, S. (2018). Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning. Information and Software Technology, 96, 94-111. https://doi.org/10.1016/j.infsof.2017.11.004

Upadhyay, S., Joshi, S., & Lakkaraju, H. (2021). Towards robust and reliable algorithmic recourse. Advances in Neural Information Processing Systems, 34, 16926-16937.

van Krieken, E., Acar, E., & van Harmelen, F. (2020). Analyzing differentiable fuzzy implications. En Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR), 2020.

van Krieken, E., Acar, E., & van Harmelen, F. (2022). Analyzing differentiable fuzzy logic operators. Artificial Intelligence, 302.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. In International Conference on Learning Representations (ICLR).

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics (TOG), 38(5), 1-12. https://doi.org/10.1145/3326362.3326440

Ward, T., & Durrant, R. (2022). Targets of explanation in correctional and forensic psychology: A black box model. Aggression and Violent Behavior, 67, 101782. https://doi.org/10.1016/j.avb.2022.101782

Wehmeier, C., & Artopoulos, G. (2023). MetaFraming: A Methodology for Democratizing Heritage Interpretation Through Wiki Surveys. In Proceedings of the 20th International Conference on Culture and Computer Science: Code and Materiality (KUI '23) (Article No. 4, pp. 1–9). https://doi.org/10.1145/3623462.3623465

Xian, T., Constantinides, P., & Mehandjiev, N. (2024). Interpretable artificial intelligence systems in medical imaging: Review and theoretical framework. En E. Pino, R. Magjarević y P. de Carvalho (Eds.), Business 2024 (pp. 240–265). DOI: https://doi.org/10.4337/9781803926216.00023

Xu, D., & Ruan, C. (Year). Modern Theoretical Tools for Understanding and Designing Next-generation Information Retrieval System. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), 1635–1637. https://doi.org/10.1145/3488560.3501394

Yik, W., Serafini, L., Lindsey, T., & Montañez, G. D. (2022). Identifying Bias in Data Using Two-Distribution Hypothesis Tests. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22) (pp. 831–844). https://doi.org/10.1145/3514094.3534169

Yordanova, Z. (2024). Ethical Implications of Transparency and Explainability of Artificial Intelligence for Managing Value-Added Tax (VAT) in Corporations. En T. Guarda, F. Portela y J.M. Diaz-Nafria (Eds.), Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2023. (Comunicaciones en Ciencia de la Computación e Información, Vol. 1936). Springer, Cham. https://doi.org/10.1007/978-3-031-48855-9_26

Zahid, F. M., Asif, I., Khan, Z. H., Shoaib, M., Nawaz, T., Liu, H., Gilman, R. H., Yousafzai, A. W., & Siddiqi, J. (2021). Implementación del aprendizaje automático para el diagnóstico automatizado del COVID-19 utilizando la arquitectura CNN. Journal of Medical Imaging and Health Informatics, 11(6), 1375-1382.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer.

Zennaro, F. M., & Ivanovska, M. (2018). Counterfactually fair prediction using multiple causal models. In European Conference on Multi-Agent Systems (pp. 249-266). Springer.

Zhang, X., Han, L., Han, L., Chen, H., Dancey, D., & Zhang, D. (2023). sMRI-PatchNet: A Novel Efficient Explainable Patch-Based Deep Learning Network for Alzheimer's Disease Diagnosis With Structural MRI. IEEE Access, 11, 108603-108616. https://doi.org/10.1109/ACCESS.2023.3321220 .

Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2017.

Zhu, K., Zhang, N., Ying, S., & Zhu, D. (2020). Within-project and cross-project just-in-time defect prediction based on denoising autoencoder and convolutional neural network. IET Software, 14(3), 185-195. https://doi.org/10.1049/ietsen.2019.0278

Zhuang, Z. Y., Churilov, L., Burstein, F., & Sikaris, K. (2009). Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. European Journal of Operational Research, 195(3), 662–675. https://doi.org/10.1016/j.ejor.2008.02.015

Zilke, J. R., Loza Mencía, E., & Janssen, F. (2016). DeepRED – Rule Extraction from Deep Neural Networks. In T. Calders, M. Ceci, & D. Malerba (Eds.), Discovery Science (pp. 457-473). Springer International Publishing. ISBN 978-3-319-46307-0 .