Alkhanbouli *et al.*
BMC Medical Informatics and Decision Making (2025) 25:110
https://doi.org/10.1186/s12911-025-02944-6

BMC Medical Informatics and Decision Making

SYSTEMATIC REVIEW

Open Access

The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions

Razan Alkhanbouli[1†], Hour Matar Abdulla Almadhaani[1†], Farah Alhosani[2] and Mecit Can Emre Simsekler[1*]

Abstract

Explainable Artificial Intelligence (XAI) enhances transparency and interpretability in AI models, which is crucial for trust and accountability in healthcare. A potential application of XAI is disease prediction using various data modalities. This study conducts a Systematic Literature Review (SLR) following the PRISMA protocol, synthesizing findings from 30 selected studies to examine XAI's evolving role in disease prediction. It explores commonly used XAI methods, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), and their impact across medical fields in disease prediction. The review highlights key gaps, including limited dataset diversity, model complexity, and reliance on single data types, emphasizing the need for greater interpretability and data integration. Addressing these issues is crucial for advancing AI in healthcare. This study contributes by outlining current challenges and potential solutions, suggesting directions for future research to develop more reliable and robust XAI methods.

Keywords Explainable artificial intelligence, XAI, Healthcare AI, Machine learning, Disease prediction, Disease recognition, Patient safety, Decision support systems, Risk management

Introduction

Artificial intelligence (AI) has been at the forefront in transforming several aspects of healthcare such as diagnosis, treatment, and disease prevention. AI can detect patterns, predict, classify, and learn from large-scale and high-dimensional data, and show remarkable performance in analyzing complex data such as medical images, multimodal physiological features, and genomic sequences. These capabilities, coupled with promising accuracies, have given AI models the potential to assist physicians and healthcare experts in making more informed decisions. However, the accuracy of the AI models often stems from an increase in model complexity hence resulting in a black-box label [1]. The black box term is used in AI to describe models that are very complex and difficult to interpret. One of the biggest challenges of AI models is that they produce outputs without stating the logic behind them. The logic then becomes hard to interpret and explain making it difficult to identify errors, biases, or inconsistencies [2]. This challenge makes it hard for healthcare professionals to trust the use of AI in healthcare settings and it also raises ethical concerns, such as accountability and responsibility of the AI models [3].

AI models, including Machine Learning (ML) and Deep Learning (DL) algorithms, might have limited transparency and interpretability in the rationale of their

[†]Razan Alkhanbouli and Hour Matar Abdulla Almadhaani contributed equally to this work and share first authorship.

*Correspondence:
Mecit Can Emre Simsekler
emre.simsekler@ku.ac.ae
[1] Department of Management Science & Engineering, Khalifa University of Science & Technology, Abu Dhabi, UAE
[2] Department of Biomedical Engineering & Biotechnology, Khalifa University of Science & Technology, Abu Dhabi, UAE

output hence the growing shift to Explainable Artificial Intelligence (XAI). There has been a notable surge in publications in XAI in the past decade [3]. XAI is a subfield of AI that incorporates transparency, interpretability, and explainability of the outcomes. Interpretable in the context of XAI is defined by Doshi-Velez and Kim as *"the ability to explain or to present in understandable terms to a human"* [1, 4]. Explainability is then defined as the understanding of the internal processes and steps taken by the model for it to come to a specific conclusion [1]. However, it is important to highlight that the explainability of the model heavily depends on the given task [3].

XAI methods enhance the understanding of AI models allowing users to have a comprehensive understanding of the strengths, limitations, and assumptions of the model. Furthermore, XAI could potentially have a significant impact on disease prediction by providing clearer insights into how AI models arrive at their conclusions, thus enabling healthcare professionals to make more informed decisions based on these predictions. Examples of XAI applications in healthcare include colorectal cancer diagnosis from histopathological images, in which important features are extracted and analyzed, and the early detection of Parkinson's Disease using DaTSCAN imagery [5, 6]. Despite the potential of XAI in facilitating decision-making in healthcare settings, the overall integration of XAI in clinical practice has been slow and limited due to the lack of trust and understanding of the models. Hence, addressing the limitations of these models could further enhance the trust and understanding of clinicians and healthcare professionals.

While there are few review articles addressing explainability, it is important to highlight that none focuses on disease prediction and recognition applications in healthcare, a significant area with potential AI implications. For instance, a recent study emphasizes comorbidity rather than the prediction of individual diseases [7]. Furthermore, previous literature reviews on XAI have covered broader applications in healthcare, such as those in the medical domain [8, 9], while others have focused on specific diseases like Alzheimer's disease [10]. Our review specifically targets XAI for disease prediction across various diseases. This focus allows us to provide a more comprehensive analysis of XAI's role in enhancing disease prediction across multiple medical conditions. Therefore, there is a gap in the literature to conduct a comprehensive review of the current literature concerning XAI methods utilized in disease prediction.

AI is increasingly transforming healthcare, especially in diagnosis, treatment planning, and disease prediction. However, most AI models operate as "black boxes," making it challenging for healthcare professionals to comprehend the decision-making process. This concerns accountability, explainability, usability and trust in essential medical circumstances [1, 3]. As a result, there is an increasing interest in Explainable AI (XAI), which aims to enhance the transparency and comprehensibility of AI decisions for medical professionals [4].

Insights into AI-driven predictions can be obtained through XAI, which will ultimately improve the safety and reliability of AI in healthcare by enabling physicians to more reliably adopt and evaluate model outputs [2]. Despite its potential, XAI is still not extensively utilized in clinical practice, especially concerning disease prediction [5]. This review focuses on applying XAI in healthcare, particularly enhancing the comprehension and applicability of AI-based disease prediction models for medical professionals. This review aims to identify and analyze the existing gaps and limitations within this domain. This systematic literature review aims to explore the existing literature on the usage of XAI methods in predicting diseases where different modalities are used, such as medical images and signals. The research questions of this study are as follows:

> Q1. What are the key XAI methods currently applied in disease prediction across different medical modalities (e.g., imaging, physiological signals)?
> Q2. What are the major limitations and challenges of existing XAI methods in enhancing transparency and interpretability in disease prediction models?

The systematic literature review is organized as follows: The theoretical background and related work on XAI in predicting diseases is included in "Theoretical background" section, highlighting previous studies and foundational concepts to frame the research context. "Research methodology" section describes the research methodology, including the search strategy, selection criteria, and data extraction and analysis methods. "Results and findings" section presents the results and findings of the literature review addressing the research questions. "Discussion and implications" section discusses the implications and limitations of the literature review as well as the directions for future research. "Gaps and solutions" section concludes the paper and summarizes the main contributions.

Theoretical background

Related work on XAI in predicting diseases

The theoretical background of existing literature on XAI in predicting disease encompasses various methodologies and applications. XAI has become increasingly vital in healthcare as AI models become more complex and integral to disease prediction and diagnosis. The overarching goal of XAI is to make AI decision-making

processes transparent and understandable, which is crucial in a field as sensitive as healthcare [7]. In disease prediction, XAI addresses the challenge of the nature of many AI models. This involves developing methods that can explain, in human-understandable terms, how AI models arrive at their conclusions. Such transparency is essential for gaining healthcare professionals' trust and adhering to regulatory standards and ethical considerations in medical practice.

The literature reveals a diverse range of XAI methods. While SHAP and LIME are prominent for their ability to provide local and global interpretability, other methods like Gradient-weighted Class Activation Mapping (Grad-CAM), Partial Dependence Plots (PDP), and Counterfactual Explanations also contribute to understanding AI decisions [11]. Each method offers unique insights, with some providing visual explanations or highlighting specific features that influence model predictions. The importance of XAI in healthcare is further underscored by its application across various diseases. From cancer detection to cardiovascular diseases, using XAI methods is instrumental in elucidating AI predictions, aiding in more accurate diagnoses and tailored treatments. This is particularly crucial in personalized medicine, where understanding the specific factors influencing a model's prediction can lead to more effective patient-specific interventions.

Among the various methodologies, LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive explanations) have emerged as prominent tools for deciphering complex models. These methods provide insights into how AI models, especially those based on DL, arrive at their predictions, thereby addressing the 'black box' nature of such models [8]. In [12], a web and mobile-based platform for diabetes diagnosis using LIME and SHAP showcases the integration of XAI into practical healthcare applications. LIME excels at providing local interpretability by approximating a model's predictions in a specific instance, but it may struggle with consistency across different datasets and can be computationally expensive for large models [3]. LIME was also used in an EEG-based machine learning model for stroke prediction, focusing on brain wave analysis, demonstrating its utility in making complex models interpretable in neurological contexts [13]. SHAP, on the other hand, provides a consistent method for calculating the contribution of each feature to the model's output, making it particularly useful for feature attribution, though it can be computationally intensive [1]. For example, SHAP has been applied to early Parkinson's Disease detection, where it highlighted significant biomarkers, aiding in the interpretability of gene expression data models [6]. Recent studies have increasingly applied both LIME and SHAP concurrently to enhance the interpretability of AI predictions in medical diagnostics. This dual approach leverages the strengths of both methods—LIME's capacity to offer local interpretability and SHAP's ability to assign consistent feature importance values across the model. Table 1 shows the objectives of the existing SLR on the XAI method.

Rather than concentrating on a particular disease, this systematic literature review distinguishes against addressing an extensive spectrum of conditions for XAI application in prediction. This comprehensive approach allows us to identify unique gaps and propose specialized solutions across various medical domains. We delve into the methodological aspects of XAI, emphasizing the distinctive opportunities and challenges involved in using these technologies for various kinds of disease predictions. Our review broadens the scope of research in the field through the integration of solutions from diverse conditions. It also provides a solid framework for addressing the identified gaps and specific needs for more targeted and effective XAI solutions in healthcare diagnostics.

Research methodology

Several health journals and databases are rich literature sources on XAI's role in predicting diseases. A systematic search was essential to identify reliable studies from credible authors, focusing on studies published in the last decade. In this context, the search prioritized studies published in the previous five years, marked by the upsurged uptake of AI technologies in healthcare to predict disease occurrence. Also, significant consideration was given to the authors' credentials, sample size, study methodologies employed, theoretical frameworks used, and thematic focus to enhance the data's relevance, validity, and reliability. This study utilized a Systematic Literature Review (SLR), which employs a transparent and rigorous approach to synthesize findings to assess and reduce bias in the data.

In this context, the review was conducted through scholarly synthesis of evidence about XAI predicting disease using critical methodology to identify, define, and assess related themes. SRV was the most preferred study design for studying XAI in predicting disease because of the significance attached to its use. For instance, the research scope requires proper canvassing of past and existing findings to unearth the strengths of XAI and other AI applications in healthcare. Specifically, SLR adopts a more scientific, reproducible, and transparent approach to data collection and analysis and offers more specific and clear guidelines to researchers in reviewing and presenting outcomes. On the same note, it is marred with fewer mistakes and biases than other study designs

Table 1 Characteristics of XAI-related studies in disease prediction (*The journal name may be abbreviated or formatted based on its indexing style. **The country information may refer to either the authors' affiliated country or the country where data was collected for the disease prediction model.)

Paper	Journal*	Country**	Disease	Data modality	Data type	Data Availability	AI Model	XAI method	Contribution
[12]	Healthcare Technology Letters	Bangladesh	Diabetes	Physiological and medical measurements	Numerical	Yes	DT SVM RF LR KNN	LIME SHAP	ML-enhanced platform for diabetes diagnosis accessible via web and mobile app
[13]	Sensors journal	Bangladesh Republic of Korea	Ischemic Stroke	Electroencephalography (EEG)	EEG Signal	No	XGBoost LGBM Adaptive Gradient Boosting	Eli5 LIME	EEG-based ML model for stroke prediction, focusing on brain wave analysis
[14]	Computational Intelligence and Neuroscience	Bangladesh Saudi Arabia	Acute Lymphocytic Leukemia (ALL)	White Blood Cells image	Image	Yes	Pretrained CNNs using Transfer Learning	LIME	InceptionV3 and LIME-based XAI model for automated Acute Lymphoblastic Leukemia detection
[15]	BMC Med Inform Decis Mak	Brazil	Sarcoidosis	Forced Oscillation Technique (FOT) exams	Numerical	Yes	KNN SVM AdaBoost RF LGBM XGBoost DT LR	Genetic programming (for intelligible expressions), Feature Importance Evaluation	Proposes a genetic programming approach with XAI for diagnosing sarcoidosis from respiratory function tests
[16]	JACC: CARDIOVASCULAR IMAGING	USA Canada Israel Switzerland	Obstructive Coronary Artery Disease (CAD)	Single-Photon Emission Computed Tomography (SPECT) Myocardial Perfusion Imaging (MPI)	Numerical	No	CAD-DL	XAI	Explainable DL model for rapid and accurate Coronary Artery Disease detection from cardiac images
[6]	Computers in Biology and Medicine	India	Parkinson's Disease (PD)	Gene Expression Data	Numerical	Yes	LR SVM DT RF KNN NB	SHAP	Applies ML and XAI to early Parkinson's Disease detection, highlighting significant biomarkers
[17]	Diagnostics	Saudi Arabia Egypt	Colon Cancer	WCE colon image dataset	image	Yes	Heterogenic stacking DL integrated with pretrained CNN models [VGG16, InceptionV3, Resnet50, DenseNet121)	XAI	Introduces a heterogenic stacking DL model with XAI for improved colon cancer prediction

Table 1 (continued)

Paper	Journal*	Country**	Disease	Data modality	Data type	Data Availability	AI Model	XAI method	Contribution
[18]	Journal of Medical Systems	China	Hepatitis	Mixture of integer and real value attributes	Numerical	Yes	LR DT KNN XGBoost SVM RF	SHAP LIME PDP	Presents an XAI framework aimed at enhancing hepatitis diagnosis and informing clinical decisions
[19]	Computer Methods and Programs in Biomedicine	South Korea USA Egypt	Parkinson's Disease (PD)	SC, BS, MH, M, and NM	Time series data	Yes	SVM RF ETC LGBM SGD	LIME SHAPASH	Proposes a multimodal time series ML pipeline with explainability for Parkinson's Disease progression prediction
[20]	IEEE Transactions on Biomedical Engineering	Algeria France	Breast Cancer	Clinicopathological data	Categorical and continuous	Yes	CatBoost	LIME	Outlines an explainable ML prognosis model for cancer metastasis, aiding personalized treatment decisions
[21]	Nature portfolio Scientific reports	Spain Egypt Republic of Korea	Alzheimer's Disease	Multimodal (PET, CSF, Cognitive Scores, Genetics, Lab Tests, Medical History, MRI, Neurological Exam and others)	Both	Yes	RF DT Binary Classification	SHAP Fuzzy	Developes a system that diagnoses and detects the progression of Alzheimer's Disease by taking into consideration different modalities
[22]	Nature research Scientific Reports	Canada	COVID-19	Chest X-ray	Image	Yes	CNN-Based Model: *Proposed COVID-Net*	GSInquire	Introduces COVID-Net which is a deep convolutional neural network that detects COVID-19 cases using explainability methods
[5]	Journal of Biomedical Informatics	Slovakia Japan	Colorectal Cancer	Histopathological data	Image	Yes	CNN	Explainable Cumulative Fuzzy Class Membership Criterion (X-CFCMC)	The classifier uses histopathological data and can predict 8 varieties of colorectal cancer

Table 1 (continued)

Paper	Journal*	Country**	Disease	Data modality	Data type	Data Availability	AI Model	XAI method	Contribution
[23]	Annals of Translational Medicine	China	Fenestral Otosclerosis	Temporal bone high-resolution computed tomography (HRCT) slices	Image	Yes	CNN-Based Model: *Proposed Otosclerosis Logical Neural Network Model*	Visualization of learned deep representations	Enhances the diagnosis of fenestral OS using temporal bone high-resolution computed tomography (HRCT) slices and DL
[24]	Computers in Biology and Medicine	India	Parkinson's Disease	Single-photon Emission Computed Tomography (SPECT) DaTSCANs	Image	Yes	VGG16	LIME	Develops an improved accuracy DL model for early diagnosis and provided visual markings generated by the model to aid medical practitioners
[25]	The Journal of Supercomputing	India	Heart Diseases	Age, Resting Blood Pressure, Exercise Induced Angina, Fasting Blood Sugar, Maximum Heart Rate, Serum Cholesterol and 7 other features	Numerical	Yes	XGBoost	SHAP LIME PDP DALEX	Works on the reduction of dimensionality using XAI while maintaining the model's accuracy
[26]	Genes	India Egypt Morocco Qatar	Cervical Cancer	Vaginal Swab Samples [Microbial Data *(16S rRNA Sequencing)]*	Genomic Sequence	Yes	RF	SHAP	Uses specific microbial patterns commonly found in cervical cancer to create personalized medicine
[27]	Cancers	Pakistan Saudi Arabia United Arab Emirates United Kingdom Saudi Arabia	Lung Pulmonary Disease	Chest Radiographs	Image	Yes	CNN-based transfer learning with RestNet50	LIME	Shows improved accuracies and explanations in interpreting pulmonary diseases using chest radiographs
[28]	Radiation Oncology	Germany	Prostate Tumor	Multi-parametric MRI	Image	Yes	U-Net architecture CNN	Grad-CAM	Proposes an XAI framework that can identify tumor prostate tissues through images

Table 1 (continued)

Paper	Journal*	Country**	Disease	Data modality	Data type	Data Availability	AI Model	XAI method	Contribution
[29]	Computers in Biology and Medicine	Nepal and Australia	COVID-19, Pneumonia, and Tuberculosis	Chest X-ray	Image	Yes	CNN	SHAP LIME Grad-CAM	Uses a CNN model to detect lung diseases using CXR images with a focus on interpretability for clinicians
[30]	Nature Communications	USA China	Kidney-related disease	Electronic health records (EHRs), medical imaging data, laboratory tests	Myocardial perfusion, wall motion, and wall thickening polar maps	No	GBT	SHAP	Creates and evaluates a transportable, XAI Model for Acute Kidney Injury Prediction
[31]	Frontiers in Medicine	Italy France	Breast cancer	Clinical and cytohistological outcomes from patients' medical records	Clinical outcomes, therapy-related information	Yes, by request	SVM RF NB XGBoost	SHAP	XAI approach reveals critical factors affecting breast cancer IDEs at 5 and 10-year post-diagnosis, aiding personalized patient care
[32]	Frontiers in Cardiovascular Medicine	Finland	Cardiovascular disease	Medical records	Numerical and nominal features	Yes	Ensemble Tree algorithm	SHAP	Creates an XAI for Heart Failure Prognosis, Balancing Clinical Insight with Predictive Precision
[33]	Frontiers in Neuroscience	Italy USA	Neuroscience	Imaging, electrophysiological recordings, clinical assessments	EEG, spike data	No	ML algorithms	LIME	Advances AI and XAI for Enhanced Brain Function and Neurostimulation Insights, Highlighting Transparent Models and Data Competitions
[34]	NPJ Digit Medicine	UK	Mental health	Clinical information, patient records, and diagnostic data	patient characteristics, diagnostic indicators, and potentially text-based information	Yes	Deep neural networks, prediction and classification in psychiatric applications	SHAP	Presents the TIFU Framework for Clear, Interpretable AI in Mental Health, with a Focus on Clinically Aligned AI Comprehensibility

Table 1 (continued)

Paper	Journal*	Country**	Disease	Data modality	Data type	Data Availability	AI Model	XAI method	Contribution
[35]	Compute Methods Programs Biomed	Spain	Prostate cancer tissue	Specifically RNAseq data obtained from transrectal biopsies	gene expression data	Yes	KNN rpart (CART) RF	SHAP	Offers a ML Classifier Using Gene Expression Data and XAI for Precise Prostate Cancer Risk Prediction
[36]	Clinical Medicine Insights Cardiology	UK and Sweden	Myocardial Infarction (MI)	Medical history, and demographic data	Physical and functional measures, and collection of blood, urine, and saliva	Yes, by request	LR XGBoost	SHAP	XGBoost with SHAP Values: A Promising Method for Predicting Myocardial Infarction Risk Across a Broad Population
[37]	Nature Communications	Denmark	Acute Critical Illness	Electronic Health Records	Secondary healthcare data from four Danish municipalities	Yes, by request	TCN	LRP	Presents an xAI-EWS System for Predicting Acute Illnesses with EHRs, Offering Clear, Real-time Insights for Clinician Decision-making
[38]	Nature Portfolio Scientific Reports I	Italy	Thyroid	Histological samples	Raman spectra	Yes, by request	ML algorithms RF XGBoost	SHAP	Combines Raman Spectroscopy and ML for Non-Invasive Thyroid Cancer Diagnosis to Potentially Lower Unneeded Surgeries
[39]	European Federation for Medical Informatics	Greece	Preterm birth	Demographics, social and medical history, and obstetrics variables	Numerical, ordinal, and nominal features	No	LR SVM RF XGBoost	SHAP	Forecasts Preterm birth chances using demographic and medical data, providing predictions and insights for enhanced pregnancy screening

because it offers high-quality evidence while leaving a transparent audit trail of the researcher's methods, inferences, and methods. Thus, it extracted data from published studies followed by analysis, description, critical appraisal, and summary interpretation into solid evidence-based conclusions.

The SLR approach facilitated a deeper understanding of the evolving landscape of XAI in healthcare. It allowed for the exploration of how different research methodologies and theoretical frameworks have influenced the development and application of XAI in disease prediction. This comprehensive review highlighted the current state of the art and identified gaps in the existing literature, suggesting areas for future research. The diverse methodologies and varied thematic focuses of the studies reviewed underscored the multifaceted nature of XAI in healthcare, pointing towards a future where AI's role in disease prediction is not only technologically advanced but also ethically sound and widely accepted in clinical practice.

A thematic approach was implemented for data collection and analysis, with research categorized according to its contribution to disease prediction, methods in which it utilized XAI techniques (e.g., SHAP, LIME), and clinical application. This method ensured that the results were accurate, consistent, and relevant to the study's objectives.

Planning the review

Researchers dedicated considerable time to a thorough brainstorming session, looking into existing methodologies for conducting systematic literature reviews. This task was achieved through extensive reading and consultation with peers engaging in similar studies to warrant uttermost adherence to the steps and processes required for conducting a systematic literature review.

During this stage, particular emphasis was placed on research ethics, especially concerning the respect for the intellectual property rights of previous researchers (proper citations to prevent plagiarism). The session succeeded by protocol review, research questions, and objectives documentation. As a best practice, the review protocol was outlined before the start to minimize the risk of unplanned research duplication and foster consistency and transparency between protocol and methodology.

Search string

A systematic literature review was performed to identify relevant studies on XAI in disease prediction. The review emphasized research published between 2014 and 2023 that indicated how AI is currently used in healthcare and how it has developed recently. A ten-year search was conducted to identify the most recent advancements in the field of XAI applications in healthcare diagnostics.

To address RQ1 and RQ2, we conducted a systematic literature review (SLR) using the PRISMA guidelines. Our search included databases such as Scopus, PubMed and Web of science, targeting studies published between 2018 and 2023 that applied XAI in disease prediction. The databases were prioritized because they comprehensively cover review papers, peer-reviewed articles, and conference proceedings. The search queries included combinations of keywords related to XAI and disease prediction, such as: "disease diagnosis" AND "machine learning," "XAI" AND "healthcare AI," and "Explainable Artificial Intelligence" AND "disease prediction." another search keywords applied were predicting disease, disease diagnosis, disease recognition, XAI in disease diagnosis, and predictive AI by adopting Boolean expressions 'OR' and 'AND'.

Using XAI techniques, articles were assessed according to the extent to which they contributed to the prediction of disease, with a focus on Grad-CAM, SHAP, LIME, and other techniques. This approach ensured that we obtained appropriate studies to investigate XAI's advantages and disadvantages. This planning phase laid a solid foundation for the literature review. It ensured the research was grounded in the latest and most relevant studies, providing a contemporary perspective on XAI in healthcare. The use of renowned databases like PubMed and Scopus guaranteed access to high-quality and peer-reviewed articles, enhancing the credibility of the research findings. Furthermore, carefully selecting search keywords and strategically using Boolean expressions enabled a comprehensive and focused literature retrieval process. This approach streamlined the review process and ensured that a broad spectrum of perspectives and findings related to XAI in disease prediction was captured for a rich and insightful analysis.

Conducting the review

Researchers thoroughly searched databases, initially identifying 76 articles. This initial selection served as a broad pool to refine and select the most relevant studies. To ensure alignment with the research objectives, each article was scrutinized based on its abstract, title, and keywords. This screening process was critical to establish each study's relevance to the review's overarching theme.

Subsequently, researchers applied specific inclusion and exclusion criteria to refine the selection of articles further. As a result of this rigorous filtering process, a total of 46 papers were excluded for various reasons. Eight of these papers were removed because they were not found in the Scopus and PubMed databases,

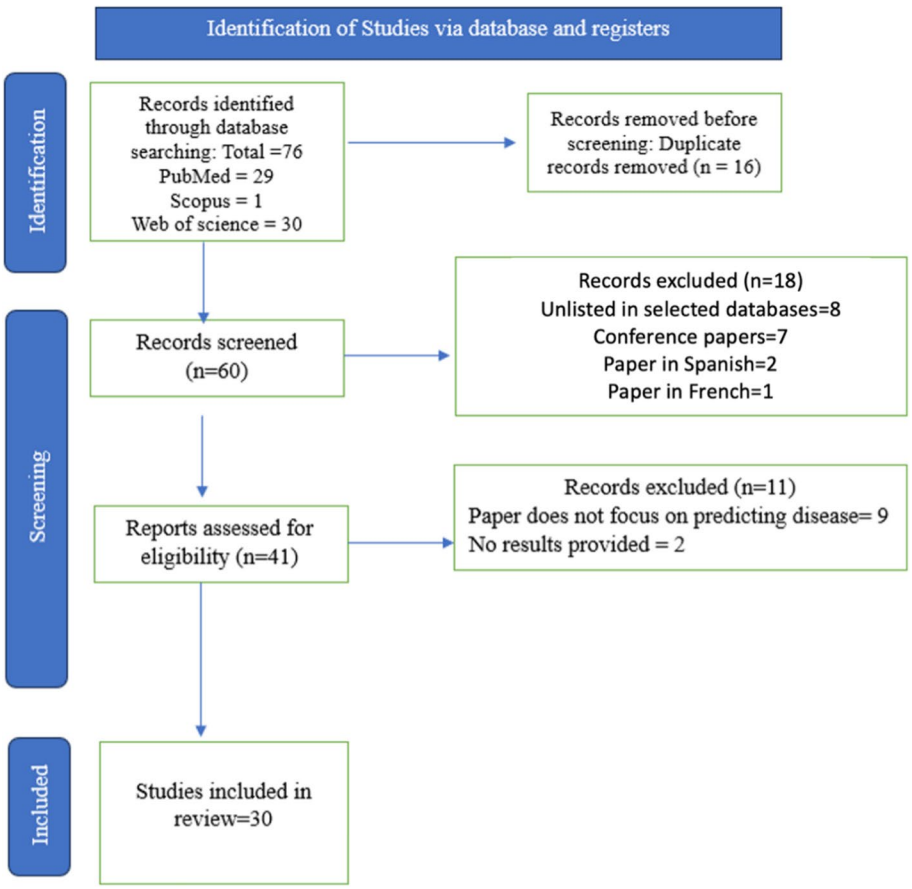

Fig. 1 Flowchart of the research methodology

indicating they might need to meet the required academic standards or relevancy. Additionally, seven papers were excluded because they were conference papers, which may have undergone a different level of peer review than journal papers. Two papers were excluded as they did not contain results. Twelve papers were deemed unrelated to the central theme of predicting disease, demonstrating the importance of thematic alignment in systematic reviews. Lastly, three papers were excluded because they were written in languages other than English, possibly hindering a thorough and accurate analysis due to language barriers. Figure 1 describes including and excluding articles during the evaluation and selection of studies.

Reporting and dissemination

This stage focused on presenting the literature review report based on the summarized findings in the results and discussion section. The collected papers were grouped and tabulated based on the paper's authors and date of publication, journal, country, disease, data modality, data type, AI model, XAI method, and contribution (see Table 1).

Results and findings

Descriptive analysis of the reviewed articles

An analysis of the papers shows that 30 articles were published across 26 different journals. This is reflected in Fig. 2, which illustrates the number of publications per journal. The journal *Computers in Biology and Medicine* published three articles, while *Nature Communications* and *Nature Portfolio Scientific Reports* published two articles each. The remaining 23 journals published one article each.

Furthermore, the number of published articles increased from 2019 to 2023, as shown in Fig. 3, with the peak in 2023 at 13 publications, followed by 9 in 2022. Only one article was published in 2019, reflecting a gradual increase from 2019 to 2023 (1 in 2019, 5 in 2020, 2 in 2021, 9 in 2022, and 13 in 2023). Overall, the years 2022 and 2023 were particularly prolific, accounting for 80% of the publications in the review period.

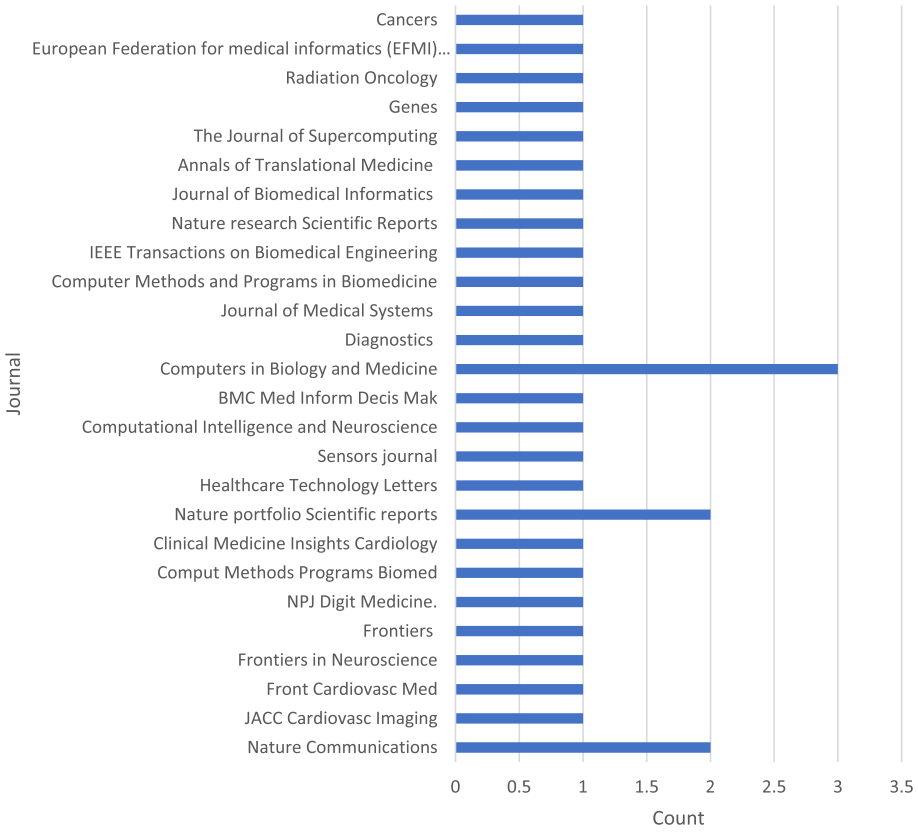

Fig. 2 Publications by journal

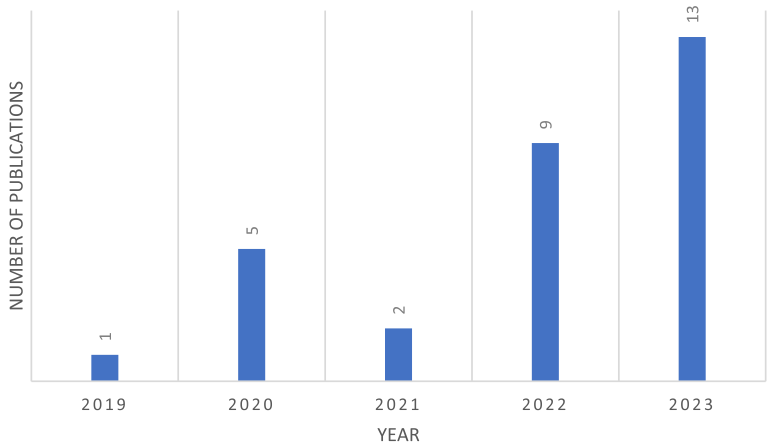

Fig. 3 Number of publications in years

The distribution of publications before 2019 was notably sparse, with only 3.33% of papers published by that time. A significant uptick in interest was observed starting in 2020, with 16.67% of the papers published that year. The subsequent years showed a fluctuating but generally increasing trend in publication volume: 10% in 2021, a notable jump to 30% in 2022, and peaking at 40% in 2023. This trend underscores a growing engagement with XAI technologies in medical diagnostics over the last few years.

Authors from 29 countries across six continents published articles related to XAI. Figure 4 shows the percentage of publications based on the country of the author, with the United States contributing the highest

Fig. 4 Articles by country of study

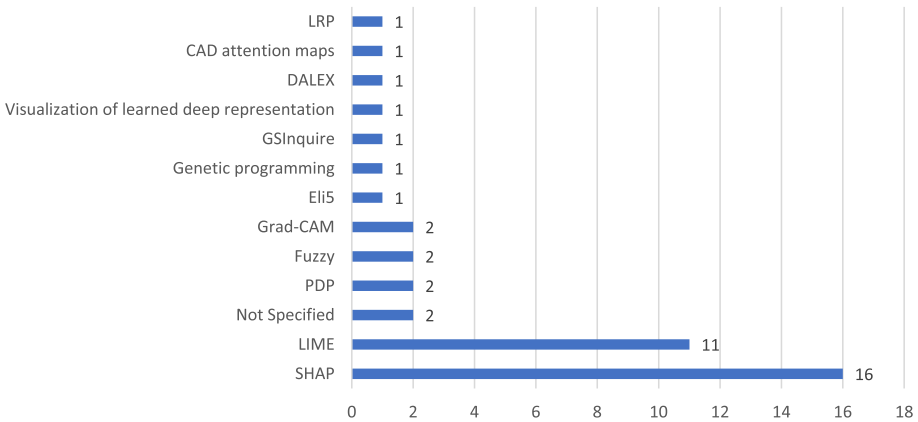

Fig. 5 XAI methods used in shortlisted papers

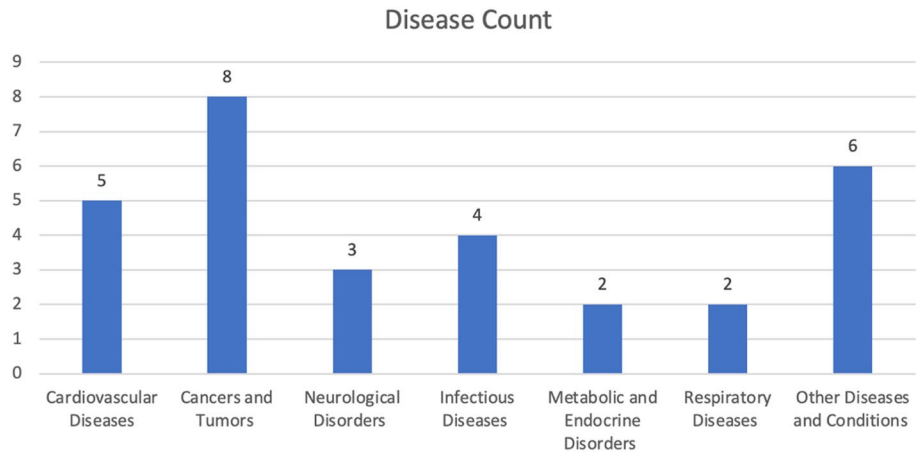

Fig. 6 Categories of diseases

at 7.5%, followed by Egypt, India, Saudi Arabia, Bangladesh, UK, Italy, and China at 6%. Most countries contributed between 2% and 5.5%, with the majority at 2%.

Findings and analysis

This section synthesizes findings from a systematic literature review (SLR) of 30 peer-reviewed articles on XAI in medical diagnostics. Figure 5 displays the various XAI methods used across the reviewed studies. SHAP was the most frequently employed method, accounting for 38% of the publications, followed by LIME at 26%. Other methods, including Grad-CAM, Fuzzy logic, and Partial Dependence Plots (PDP), were each used in 5% of the studies. Additionally, a few methods—such as Eli5, Genetic Programming, GSInquire, Visualization of learned deep representations, CAD attention maps, Layer-wise Relevance Propagation (LRP), and DALEX were employed minimally, representing only 1% of the publications each.

Figure 5 highlights the dominant roles of SHAP and LIME in improving AI model interpretability for conditions like Diabetes, Ischemic Stroke, and various cancers. Moreover, the surveyed literature indicates an exploration of alternative XAI techniques, which contribute to the refinement of diagnostic precision. Notably, genetic programming, as reported in [15] has shown promise in the diagnosis of sarcoidosis, while the implementation of Fuzzy logic has been instrumental in evaluating Alzheimer's Disease, as discussed in [21].

Table 1 presents a comprehensive overview of AI models applied to disease diagnosis across different data modalities, including EEG signals for Ischemic Stroke, chest X-rays for COVID-19, and gene expression profiling for Prostate Cancer. Each study, such as [12] for diabetes diagnosis [16], predictive model for Coronary Artery Disease, and [26] prognostic method for cervical cancer, showcases AI's adaptability in healthcare.

Figure 6 categorizes the diseases studied into seven principal groups: cardiovascular, cancers and tumors, neurological, infectious, metabolic and endocrine, respiratory, and other conditions. Cardiovascular diseases include Ischemic Stroke, Coronary Artery Disease, and Myocardial Infarction, while cancers cover conditions like Leukemia, Breast, Colon, and Prostate cancers. Neurological disorders such as Parkinson's and Alzheimer's are also prominent. Infectious diseases include COVID-19, Pneumonia, Tuberculosis, and Hepatitis. Other diseases such as Sarcoidosis and Acute Critical Illness further underscore the broad scope of XAI's application.

This synthesis, as shown in Fig. 6, showcases both the complexity inherent in medical research and the expansive potential of XAI in addressing diverse diagnostic challenges. It reflects the wide range of health issues impacting humans, emphasizing the need for adaptable and innovative approaches in health sciences. This comprehensive categorization highlights the significant role XAI can play in transforming diagnostics across a broad spectrum of medical domains, enhancing the accuracy and interpretability of AI-driven solutions in healthcare.

According to our review, the most widely used XAI algorithms for disease prediction, particularly for diagnosing cancer and cardiovascular diseases, are SHAP (38%) and LIME (26%). While LIME has been applied in neurological and cancer studies, including imaging data, SHAP has often been used in gene expression studies for diseases like Parkinson's.

Our review has identified several sets of limitations in the reviewed papers. One of the most significant problems is the lack of different datasets—most models

depend on single-modality data, which reduces the conclusions' generalizability. Furthermore, clinicians have reported challenges in understanding the outcomes of XAI models [40]. This underscores the urgent need for more approachable tools and improved instruction in XAI techniques.

The review discovered a growing trend of XAI integration with multi-modal inputs such as genetic, physiological, and imaging data, as well as complicated medical datasets. In order assist with personalised treatment decisions, a study on prostate cancer that used SHAP demonstrated how the approach may highlight significant components from patient data. The findings did, however, also indicate a variety of gaps in the literature, including the lack of diverse datasets and the need for improved user interfaces so that physicians can successfully understand AI outputs.

Discussion and implications

The descriptive analysis of the reviewed papers reveals a significant and growing interest in XAI within the medical field, as demonstrated by the steady increase in publications from 2019 to 2023. This surge underscores the healthcare sector's pressing need for transparent AI systems, particularly for disease prediction and diagnosis. The global recognition of XAI's importance is evident from the diverse range of journals and the wide geographical spread of authors, with contributions spanning Asia, Europe, and the United States. This suggests that XAI is becoming an essential tool across different healthcare systems, with SHAP and LIME emerging as the most prominent methods, used in over half of the studies reviewed. Their effectiveness in enhancing AI model interpretability is well-established, but the continued exploration of other methods reflects the evolving nature of XAI, highlighting room for innovation and improvement in achieving optimal explainability.

The concentration of publications in recent years, particularly in 2022 and 2023, signals a rapid response to the growing complexity of AI technologies in healthcare. This shift indicates that the field is maturing, but it also points to the need for further refinement and development of XAI models to ensure their practical application in clinical settings. The geographical distribution of research highlights how healthcare systems globally are acknowledging the potential of XAI, yet it also emphasizes the disparities in research outputs between regions, with Asia and Europe leading, while other continents remain underrepresented. This calls for more collaborative efforts to ensure equitable access to XAI advancements across all regions.

The implications of these findings are significant. The increasing volume of literature reflects a readiness within the healthcare sector for wider adoption of XAI technologies, but challenges remain. There is a clear need for standardization in XAI methodologies and the development of comprehensive guidelines to facilitate their integration into clinical workflows. Moreover, as XAI becomes more embedded in healthcare, educational initiatives must be implemented to equip clinicians with the skills to interpret AI-assisted diagnostics effectively [7]. The variety of XAI methods utilized across studies suggests that the field is poised for further research and development. Future innovations could focus on improving model accuracy, reducing computational overhead, automation bias, and targeting specific medical conditions to enhance diagnostic precision [41].

As AI continues to play an increasing role in healthcare, the establishment of robust policy frameworks and regulatory guidelines will be critical to ensuring the ethical and safe deployment of XAI technologies. Without clear standards, the integration of these tools could face significant barriers. The findings from this review highlight the urgent need for ongoing research, regulatory oversight, and cross-disciplinary collaboration to advance XAI's role in transforming healthcare delivery.

Gaps and solutions

The systematic literature review conducted on the application of XAI in medical diagnostics has revealed several gaps in current methodologies and their implementation. These gaps span the scope of XAI models, their interpretability in clinical practice, the technologies underpinning them, and their integration into healthcare systems. As illustrated in Table 2, our findings indicate the necessity for XAI models to align closely with medical diagnostic processes, ensuring that the outputs are transparent and interpretable for medical professionals. These models must consider the complex nature of medical data and elucidate reasoning that builds clinician trust and supports clinical decision-making.

To address the identified gaps, our review recommends a comprehensive strategy. Assembling diverse datasets is essential to reduce bias and enhance the generalizability of XAI in healthcare. This includes global partnerships for data collection and techniques like SMOTE/ADASYN for balancing datasets. Additionally, achieving a balance between model complexity and interpretability is critical. Methods such as SHAP and LIME can enhance user engagement and improve understanding of the algorithms. From a technological standpoint, deploying XAI effectively relies on big data analytics to uncover deeper insights from complex medical data. This approach can improve diagnostic accuracy and patient outcomes.

Table 2 Gaps and potential solutions for XAI applications in disease prediction

	Gaps	Solutions
Model Scope	Limited datasets affecting model diversity and bias	Partner globally for diverse datasets; use synthetic data to mitigate bias
	Data imbalance skewing predictive outcomes	Employ SMOTE/ADASYN techniques for balanced datasets
	Inadequate development of explainable, transparent models	Adopt XAI frameworks, conduct audits, and provide training for healthcare providers
Modeling Approach	Struggle to balance complex models with user interpretability	Use SHAP and LIME for interpretability
	Dependency on single data types of limits prediction scope	Support interdisciplinary innovation for data integration
	Narrow performance metrics focus, overlooking comprehensive assessment	Tailor metrics to clinical outcomes and provider needs
Technology	"Black box" models obscure operational understanding	Build transparent XAI models
	Single-data modality fails to offer a complete diagnostic picture	Create simulation tools for single-modal data insights
	AI interpretability not aligned with clinical reasoning	Use AI coaching to enhance clinical reasoning
Implementation	AI interfaces lack accessibility for medical staff	Design user-centered AI interfaces with customizable options
	Complex AI tools challenge clinical workflow integration	Create modular AI tools for seamless workflow integration and training

Simulation tools should also be developed for deeper insights using single-modal data.

For successful implementation, creating user-friendly AI interfaces is vital. These interfaces should be designed with a user-centered approach, providing clear explanations of AI's reasoning. AI 'coaching' systems can further enhance healthcare providers' clinical reasoning. By implementing these strategies, the potential of XAI in medical diagnostics ensure that AI applications are transparent, verifiable, and fully integrated into patient care, thus enhancing the overall quality and efficiency of healthcare delivery.

Conclusion

In conclusion, this systematic literature review has provided a detailed analysis of the current state of XAI in disease prediction within the healthcare sector. The review highlights the growing adoption of methodologies such as SHAP and LIME, which play a crucial role in enhancing the interpretability and transparency of AI models used in complex medical diagnostics.

Our findings indicate a significant increase in relevant publications from 2019 to 2023, reflecting the rising demand for explainable AI systems in healthcare. This trend underscores the importance of XAI in supporting more informed and accountable decision-making in medical practice. Despite these advancements, challenges remain—particularly regarding the need for more diverse and comprehensive datasets to improve the generalizability and fairness of XAI models. The trade-off between model complexity and interpretability also persists, requiring ongoing efforts to develop AI systems that are both sophisticated and accessible to healthcare practitioners. While XAI is rapidly gaining traction in the healthcare domain, its full potential has yet to be realized. Future research should focus on closing these gaps by exploring more diverse data sources, simplifying the complexity of AI models, and ensuring their practical integration into clinical settings without compromising performance. Additionally, continued evaluation of XAI's real-world applications will be crucial in determining its long-term impact on healthcare.

This review contributes to a better understanding of the current landscape of XAI in healthcare, offering a foundation for future studies and guiding healthcare professionals and AI developers in the responsible and effective implementation of AI technologies in disease prediction and diagnosis. It aims to guide healthcare professionals and AI developers toward responsible and effective implementation of AI technologies, ultimately enhancing the quality and efficiency of healthcare services.

One notable limitation of this review is the exclusion of large language model (LLM)-based explanation methods, which have gained traction in recent years. Given that our study is based on literature published within a specific timeframe, many advancements in LLM-driven explainability were not yet reflected in our dataset. Future research should explore the role of LLMs in enhancing interpretability, particularly in multimodal healthcare applications, to assess their impact on trust, usability, and clinical decision-making.

Acknowledgements
NA.

Authors' contributions
R.A., H.A. and F.A. wrote the main manuscript text and prepared all figures and tables. M.S. validated the results and revised the manuscript. All authors reviewed and approved the manuscript.

Funding
Not applicable.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 January 2025 Accepted: 20 February 2025
Published online: 04 March 2025

References

1. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: a review of machine learning interpretability methods. Entropy. 2021;23:1–45. MDPI AG.
2. Aldoseri A, Al-Khalifa KN, Hamouda AS. Re-think data strategy and integration for artificial intelli-gence: concepts, opportunities and challenges. 2023; Available from: www.preprints.org.
3. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors. 2023;23:634. MDPI.
4. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017; Available from: http://arxiv.org/abs/1702.08608.
5. Sabol P, Sinčák P, Hartono P, Kočan P, Benetinová Z, Blichárová A, et al. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. J Biomed Inform. 2020;1:109.
6. Bhandari N, Walambe R, Kotecha K, Kaliya M. Integrative gene expression analysis for the diagnosis of Parkinson's disease using machine learning and explainable AI. Comput Biol Med. 2023;1:163.
7. Alsaleh MM, Allery F, Choi JW, Hama T, McQuillin A, Wu H, et al. Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: a systematic review. Int J Med Inform. 2023;175:105088. Elsevier Ireland Ltd.
8. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. Comput Biol Med. 2023;166:107555. Elsevier Ltd.
9. Sreeja MU, Philip AO, Supriya MH. Towards explainability in artificial intelligence frameworks for heartcare: a comprehensive survey. J King Saud Univ Comput Inf Sci. 2024;36:102096. King Saud bin Abdulaziz University.
10. Vimbi V, Shaffi N, Mahmud M. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. Brain Inform. 2024;11:10. Springer Science and Business Media Deutschland GmbH.
11. Elkhawaga G, Elzeki O, Abuelkheir M, Reichert M. Evaluating explainable artificial intelligence methods based on feature elimination: a functionality-grounded approach. Electronics (Switzerland). 2023;12(7):1670.
12. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2023;10(1–2):1–10.
13. Islam MS, Hussain I, Rahman MM, Park SJ, Hossain MA. Explainable artificial intelligence model for stroke prediction using EEG signal. Sensors. 2022;22(24):9859.
14. Abir WH, Uddin MF, Khanam FR, Tazin T, Khan MM, Masud M, et al. Explainable AI in diagnosing and anticipating leukemia using transfer learning method. Comput Intell Neurosci. 2022;2022:5140148.
15. de Lima AD, Lopes AJ, do Amaral JLM, de Melo PL. Explainable machine learning methods and respiratory oscillometry for the diagnosis of respiratory abnormalities in sarcoidosis. BMC Med Inform Decis Mak. 2022;22(1):274.
16. Otaki Y, Singh A, Kavanagh P, Miller RJH, Parekh T, Tamarappoo BK, et al. Clinical deployment of explainable artificial intelligence of SPECT for diagnosis of coronary artery disease. JACC Cardiovasc Imaging. 2022;15(6):1091–102.
17. Gabralla LA, Hussien AM, AlMohimeed A, Saleh H, Alsekait DM, El-Sappagh S, et al. Automated diagnosis for colon cancer diseases using stacking transformer models and explainable artificial intelligence. Diagnostics. 2023;13(18):2939.
18. Peng J, Zou K, Zhou M, Teng Y, Zhu X, Zhang F, et al. An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. J Med Syst. 2021;45(5):61.
19. Junaid M, Ali S, Eid F, El-Sappagh S, Abuhmed T. Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease. Comput Methods Programs Biomed. 2023;1:234.
20. Maouche I, Terrissa LS, Benmohammed K, Zerhouni N. An explainable AI approach for breast cancer metastasis prediction based on clinicopathological data. IEEE Trans Biomed Eng. 2023;70:3321–9.
21. El-Sappagh S, Alonso JM, Islam SMR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep. 2021;11(1):2660.
22. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep. 2020;10(1):19549.
23. Tan W, Guan P, Wu L, Chen H, Li J, Ling Y, et al. The use of explainable artificial intelligence to explore types of fenestral otosclerosis misdiagnosed when using temporal bone high-resolution computed tomography. Ann Transl Med. 2021;9(12):969–969.
24. Magesh PR, Myloth RD, Tom RJ. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. Comput Biol Med. 2020;1:126.
25. Das S, Sultana M, Bhattacharya S, Sengupta D, De D. XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI. J Supercomput. 2023;79(16):18167–97.
26. Sekaran K, Varghese RP, Gopikrishnan M, Alsamman AM, El Allali A, Zayed H, et al. Unraveling the dysbiosis of vaginal microbiome to understand cervical cancer disease etiology—an explainable AI approach. Genes (Basel). 2023;14(4):936.
27. Naz Z, Khan MUG, Saba T, Rehman A, Nobanee H, Bahaj SA. An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs. Cancers (Basel). 2023;15(1):314.
28. Gunashekar DD, Bielak L, Hägele L, Oerther B, Benndorf M, Grosu AL, et al. Explainable AI for CNN-based prostate tumor segmentation in multiparametric MRI correlated to whole mount histopathology. Radiat Oncol. 2022;17(1):65.
29. Bhandari M, Shahi TB, Siku B, Neupane A. Explanatory classification of CXR images into COVID-19, pneumonia and tuberculosis using deep learning and XAI. Comput Biol Med. 2022;1:150.
30. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun. 2020;11(1):5668.
31. Massafra R, Fanizzi A, Amoroso N, et al. Analyzing breast cancer invasive disease event classification through explainable artificial intelligence. Front Med. 2023;10:1116354. https://doi.org/10.3389/fmed.2023.1116354.
32. Moreno-Sánchez PA. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. Front Cardiovasc Med. 2023;10:1219586.
33. Fellous JM, Sapiro G, Rossi A, Mayberg H, Ferrante M. Explainable artificial intelligence for neuroscience: behavioral neurostimulation. Front Neurosci. 2019;13:13.
34. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med. 2023;6:6. Nature Research.
35. Ramírez-Mena A, Andrés-León E, Alvarez-Cubero MJ, Anguita-Ruiz A, Martinez-Gonzalez LJ, Alcala-Fdez J. Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. Comput Methods Programs Biomed. 2023;1:240.

36. Moore A, Bell M. XGBoost, A Novel Explainable AI Technique, in the prediction of myocardial infarction: a UK Biobank cohort study. Clin Med Insights Cardiol. 2022;16:11795468221133612.
37. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun. 2020;11(1):3852.
38. Bellantuono L, Tommasi R, Pantaleo E, Verri M, Amoroso N, Crucitti P, et al. An eXplainable artificial intelligence analysis of Raman spectra for thyroid cancer diagnosis. Sci Rep. 2023;13:16590.
39. Kyparissidis Kokkinidis I, Logaras E, Rigas ES, Tsakiridis I, Dagklis T, Billis A, et al. Towards an explainable ai-based tool to predict preterm birth. Stud Health Technol Inform. 2023;18;302:571–5.
40. Al-Absi DT, Simsekler MCE, Omar MA, et al. Exploring the role of artificial intelligence in acute kidney injury management: a comprehensive review and future research agenda. BMC Med Inform Decis Mak. 2024;24:337.
41. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil. 2024;5(4):460–9.

Publisher's Note