

The application of explainable artificial intelligence in the prediction, diagnoses, treatment, and management of chronic diseases: A systematic review

Hooman Hoghooghi Esfahani¹, Shogo Toyonaga² and Kiemute Oyibo² 

Abstract

Study objectives: This systematic review analyzes the applications of explainable artificial intelligence (XAI) algorithms in chronic disease care, focusing on prediction, diagnosis, treatment, and management. The study examines prevalent XAI approaches across different chronic conditions and evaluates research gaps.

Methods: The review followed Preferred Reporting Items for Systematic Review and Meta-analysis 2020 guidelines, analyzing relevant articles from 6 databases to identify and evaluate XAI implementations in chronic disease care. A protocol for this systematic review was not registered anywhere prior to publication.

Results: Three primary XAI techniques emerged as dominant: SHapley Additive exPlanations (SHAP) (46.5%), Local Interpretable Model-Agnostic Explanations (25.8%), and Gradient-weighted Class Activation Mapping (Grad-CAM) (12.0%). Disease prediction dominated the applications (86.2%), with SHAP being preferred for structured clinical data and Grad-CAM showing strength in medical imaging. Implementation varied significantly across different chronic conditions, with standardized diagnostic criteria and structured data receiving more attention.

Discussion: The analysis revealed an imbalance in healthcare applications, with sophisticated prediction models but limited treatment planning and disease management implementations. Key challenges included insufficient handling of complex multimodal data types and limited data volume. The need for extensive clinical validation in real-world settings was identified as crucial for establishing practical utility.

Conclusion: While XAI shows promise in chronic disease healthcare, advancement requires expanding beyond prediction into treatment and management domains, developing robust approaches for complex medical data, and implementing larger-scale studies. Success depends on collaboration between AI researchers, healthcare professionals, legal experts, and policymakers, alongside clear regulatory guidelines and governance frameworks balancing innovation with patient privacy.

Keywords

Explainable artificial intelligence, chronic diseases, explainable artificial intelligence algorithms, machine learning in healthcare

Received: 25 March 2025; accepted: 17 June 2025

Introduction

Rationale

Chronic diseases, also known as noncommunicable diseases (NCDs), are long-lasting conditions that persist for one or more years and require constant medical attention. Furthermore, they are defined as conditions that necessitate

¹Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

²Department of Electrical Engineering and Computer Science, York University, Toronto, Ontario, Canada

Corresponding author:

Kiemute Oyibo, Department of Electrical Engineering and Computer Science, York University, Toronto, Ontario, Canada M3J 1P3.
Email: kiemute.oyibo@yorku.ca



ongoing care, limit daily activities, or both. These diseases are characterized by their extended duration and generally slow progression.¹

According to the World Health Organization (WHO), chronic diseases are the leading cause of death and disability worldwide.^{2–4} The most common chronic diseases globally include the following:

1. Cardiovascular diseases (e.g., heart disease and stroke)
2. Cancer
3. Chronic respiratory diseases (e.g., chronic obstructive pulmonary disease and asthma)
4. Diabetes⁴

Statistics according to the WHO, show that NCDs are a significant global health challenge, accounting for 71% of all deaths worldwide, approximately 41 million people each year. Among NCDs, cardiovascular diseases are the leading cause of mortality, responsible for 17.9 million deaths annually. Cancer is the second leading cause of death globally, with 9.3 million fatalities recorded in 2020. Additionally, chronic respiratory diseases and diabetes contribute significantly to the global burden of NCDs, causing 4.1 million and 1.5 million deaths, respectively in 2019. Chronic NCDs are a significant cause of premature death worldwide, leading to substantial economic losses. For instance, in New Zealand, the economic consequences of physical inactivity—an important risk factor for NCDs—are considerable, affecting both healthcare costs and overall economic output.⁵

The evolution of chronic disease prediction and management has transformed significantly over the past decades. Before the advent of machine learning, healthcare providers relied primarily on statistical risk scores and clinical guidelines, such as the Framingham Risk Score (introduced in 1948) for cardiovascular disease, which achieved accuracy rates of 60–70%.⁶

Explainable artificial intelligence (XAI) refers to methods and techniques in the application of artificial intelligence (AI) such that the results of the solution can be understood by humans.⁷

The use of XAI in healthcare is crucial due to several factors as follows:

1. Transparency: XAI provides clarity on how AI models make decisions, which is essential in medical contexts where the rationale behind a diagnosis or treatment recommendation is critical.⁸
2. Trust: By making AI decision-making processes more understandable, XAI can help build trust among healthcare providers and patients in AI-assisted healthcare.⁹
3. Regulatory compliance: Many regulatory bodies require explainability in AI systems used in

healthcare to ensure patient safety and ethical use of technology.¹⁰

4. Improved decision making: XAI can help healthcare professionals make more informed decisions by providing interpretable insights from complex data.¹¹

Some common XAI techniques used in healthcare include the following:

1. Local Interpretable Model-agnostic Explanations (LIME)¹²
2. Shapley Additive exPlanations (SHAP)¹³
3. Counterfactual Explanations¹⁴
4. Attention Mechanisms in Deep Learning Models¹⁵

The application of XAI in chronic disease management offers several potential benefits.

Early detection

XAI significantly improves early detection of chronic diseases by analyzing patterns and risk factors. For instance, a study on AI-powered wearables demonstrated that these devices can continuously monitor vital signs such as heart rate and glucose levels, enabling timely interventions. The continuous data collected allows healthcare professionals to detect abnormalities early, potentially reducing hospital readmissions by up to 30% for patients with chronic conditions like diabetes and cardiovascular diseases.¹⁶

Personalized treatment

Personalized treatment is another critical area where XAI excels. A nationwide chronic-disease management solution implemented in Turkey utilized clinical decision support (CDS) services to tailor treatment plans based on individual patient data. This approach led to a 25% increase in adherence to evidence-based clinical guidelines, showcasing the effectiveness of personalized interventions in improving health outcomes for chronic disease patients.¹⁷ Moreover, studies on personalized medicine have shown that targeted treatments can improve patient outcomes by as much as 40% compared to standard treatment protocols.¹⁸

Patient engagement

XAI enhances patient engagement by making complex medical information accessible and understandable. Research indicated that effective communication within interprofessional teams significantly boosts patient engagement in chronic disease management programs. A qualitative study found that 85% of patients reported improved understanding of their health conditions when provided with clear explanations from healthcare providers using XAI tools.¹⁹ Additionally, the integration of AI-driven

wearables has been shown to empower patients actively, leading to a 20% increase in self-management behaviors among users.¹⁶

Clinical decision support

XAI provides valuable support for clinical decision-making by delivering interpretable insights that inform diagnosis and treatment strategies. A pilot study evaluating the Nutri CDS System found that 100% of participating clinicians were able to set high-impact diet goals during collaborative discussions with patients, which indicates the system's effectiveness in enhancing clinical workflows.²⁰ Furthermore, implementing CDS systems in chronic disease management has been associated with a 15% improvement in the quality of care delivered, as evidenced by adherence to clinical guidelines and patient satisfaction scores.²¹

Objectives

The primary objective of this systematic review is to comprehensively examine and synthesize the current state of research on the application of XAI in the context of chronic diseases. This review aims to provide a thorough understanding of how XAI technologies are being utilized across various aspects of chronic disease management, from prediction and diagnosis to treatment and long-term care.

The research questions guiding this systematic review are as follows:

RQ1. How has XAI been applied in the prediction, diagnosis, treatment, and management of chronic diseases?

This overarching question encompasses several key areas of investigation such as the following:

- (a) **Prediction:** How are XAI methods being used to predict the onset or progression of chronic diseases? This includes risk assessment models and early detection systems that provide interpretable results.
- (b) **Diagnosis:** What role does XAI play in improving the accuracy and transparency of diagnostic processes for chronic conditions? This involves examining how XAI techniques enhance the interpretation of complex medical data, such as imaging results or multi-factorial diagnostic criteria.
- (c) **Treatment:** How are XAI approaches contributing to treatment planning and decision-making in chronic disease care? This includes the use of XAI in personalized medicine approaches and in explaining treatment recommendations to both healthcare providers and patients.
- (d) **Management:** What is the impact of XAI on the long-term management of chronic diseases? This encompasses the use of XAI in monitoring disease

progression, predicting exacerbations, and supporting patient self-management.

RQ2. How has XAI been integrated across different chronic disease domains, and which areas demonstrate the highest concentration of XAI applications?

Through this research question, we aim to systematically analyze the distribution and implementation of XAI methods across major chronic disease domains, including cardiovascular diseases, diabetes, chronic respiratory diseases, chronic kidney diseases (CKDs), cancers, and neurological disorders. Our systematic review will examine how XAI has been integrated into diagnostic, treatment, and management processes within each disease domain. This analysis will reveal the current landscape of XAI adoption, identifying areas of concentrated development and potential gaps in implementation. By mapping the distribution of XAI applications across these chronic conditions, we will provide insights into which disease domains have seen the most extensive integration of XAI technologies. This comprehensive assessment will contribute to understanding the current state of XAI in chronic disease care and illuminate opportunities for future research and development in underserved areas.

RQ3. What are the emerging patterns and most successful machine learning approaches in developing interpretable AI solutions for chronic condition care?

With our third research question, we aim to identify and analyze the current trends and most successful machine learning models employed in XAI systems specifically designed for chronic disease management. Our objective is to provide a comprehensive overview of the AI landscape in this domain, highlighting which models and techniques are gaining traction and demonstrating effectiveness. We will examine various machine learning approaches, including but not limited to decision trees, random forests (RFs), neural networks, and ensemble methods, assessing their prevalence and efficacy in creating explainable outcomes. By identifying prevailing trends, we seek to understand the direction in which the field is moving and what factors are driving these trends. Furthermore, by evaluating the effectiveness of different models, we aim to discern which approaches are yielding the most promising results in terms of accuracy, interpretability, and clinical utility.

RQ4. Which XAI techniques are predominantly employed across the spectrum of chronic disease care—from prediction and diagnosis to treatment and management?

This research question aims to identify and analyze the most frequently utilized XAI methods across the entire life-cycle of chronic disease management. Our objective is to provide a comprehensive overview of the XAI landscape in chronic care, highlighting which techniques are gaining traction in different stages: prediction, diagnosis, treatment planning, and long-term management. We seek to understand not only which methods are prevalent but also how

they contribute to making complex AI models more interpretable and trustworthy in healthcare settings. This analysis will involve examining various XAI approaches, such as LIME, SHAP and Gradient-weighted Class Activation Mapping (Grad-CAM), assessing their applicability and effectiveness in different chronic disease contexts.

By addressing these aspects, this systematic review aims to accomplish the following:

1. Identify and categorize the various XAI techniques currently being applied in chronic disease care.
2. Evaluate the effectiveness and limitations of these XAI applications across different chronic conditions.
3. Identify current research gaps, challenges, and propose future directions in the application of XAI for chronic disease care.

There are other studies such as,²² which examined XAI in the broader context of healthcare CDS systems. Some other research works have also focused on the application of XAI in addressing specific chronic diseases such as Arrieta et al.⁷ in prediction of diabetes, Guleria et al.²³ in prediction of cardiovascular disease and Sethi et al.²⁴ in prediction of heart disease. Some other research works addressed the diagnoses of chronic diseases such as Jagadeesh²⁵ and Moreno-Sánchez²⁶ in early diagnosis of CKD.

In contrast to the above papers which mainly focus on a specific chronic disease, our paper is a comprehensive study focusing on the XAI applications prediction, diagnoses, treatment, and management of different chronic diseases.

This approach allows us to delve deeper into the unique challenges and opportunities presented by chronic diseases, which often require long-term care, continuous monitoring, and personalized treatment plans. By concentrating on chronic diseases, we aim to uncover specific XAI applications that address the complexities of managing conditions like diabetes, cardiovascular diseases, chronic respiratory diseases, and cancer. This targeted approach will provide valuable insights for healthcare providers, researchers, and technology developers working specifically in the field of chronic disease management.

Furthermore, our study will explore how XAI methods are being used across the entire spectrum of chronic disease care, from early prediction and diagnosis to long-term management and treatment optimization. This comprehensive view will help identify areas where XAI has made significant impacts in chronic disease care, as well as areas that require further development or research attention.

Methods

Eligibility criteria

The inclusion and exclusion criteria are outlined in Table 1. Firstly, included articles were required to be peer-reviewed journal or conference papers written in English. Secondly,

Table 1. Eligibility criteria for included and excluded articles from the systematic review.

Inclusion criterion	Exclusion criterion
Article must be a peer-reviewed paper from a journal or conference proceeding.	Article is not a book chapter, magazine write-up, online essay, etc.
Article is written in English.	Article is written in other languages.
Article is based on primary data.	Article is based on secondary data.
Article focuses on the use of XAI to treat, manage, or predict a chronic illness.	Article focuses on the development and deployment of a black-box machine learning model without discussing explainability.

the articles must focus on the direct usage of XAI to treat, manage, or predict a chronic illness. Chronic illness is defined as a persistent, long-lasting (three months or more), which negatively affects a person's quality of life (QoL). Lastly, the articles must be based on primary data.

Information sources and search strategy

Eight databases (ACM Digital Library, CINAHL, Institute of Electrical and Electronics Engineers (IEEE) Xplore, Medline, Public Health Database, PubMed, SCOPUS, Web of Science) were searched on May 22, 2024. The set of keywords used for searching the eight databases included **“explainable AI” OR “explainable artificial intelligence” OR XAI) AND chronic AND (disease* OR disorder* OR illness* OR sickness* OR infection* OR condition*)**.

Selection and data collection process

The first and second authors independently screened 111 unique articles obtained from the database search based on title, abstract, and keyword screening which resulted in 73 articles for full-text review. Articles with an unclear decision were automatically moved into the full-text review. Both authors independently completed the full-text review and compared results.

The full-text review process resulted in 58 articles to include in the final analysis. Collaborative discussions regarding discrepancies from the independent full-text review were not required as the interrater reliability score between the two authors was 1.0. The primary criterion for inclusion of an article apart from being peer-reviewed and written in English was that it uses machine learning

techniques in the treatment, management, or prediction of chronic illnesses while simultaneously being explainable to elucidate the black-box techniques.

Data items

Data was charted independently and synthesized in Google Sheets and Draw.io. Data items were chosen in accordance with the objectives of the systematic review. Table 2 outlines the critical data items that were extracted from each included article with their corresponding definition.

Study risk of bias assessment

The risk of bias (RoB) for each study was assessed using Cochrane's assessment tool.²⁷ The assessment of bias impact considered factors such as participant characteristics (including sample size and demographic information), the presence or absence of control or comparison groups, and the specific evaluation metrics used to determine study outcomes. Eventually, five categories of RoB (detailed in Table 3) for each of the included articles are specified.

Effect measures

The outcomes of each study were assessed by the first two authors independently and collaboratively. In this systematic review, a range of classification metrics to evaluate the performance of machine learning algorithms integrated with various XAI methods in the context of chronic disease management are employed. The primary metrics utilized were accuracy, precision, recall (sensitivity), F1-score, specificity, and the area under the receiver operating characteristic curve (AUC-ROC). Accuracy provided an overall measure of correct predictions, while precision and recall offered insights into the models' ability to correctly identify positive cases and avoid false negatives, respectively. The F1-score, being the harmonic mean of precision and recall, served as a balanced measure of the model's performance. Specificity was included to assess the algorithms' capability in correctly identifying negative cases. Lastly, the AUC-ROC was employed to evaluate the models' discriminative ability across different classification thresholds. These metrics, when considered collectively, offered a comprehensive assessment of the machine learning algorithms' performance while also allowing for meaningful comparisons between different XAI approaches in terms of their impact on model interpretability and predictive power in chronic disease scenarios. Moreover, there were no studies that used Qualitative reporting measures.

Synthesis methods

Articles resulting from the full-text review were classified into groups based on the following evaluation metrics:

Table 2. Article and XAI-based intervention characteristics and descriptions.

Data item	Description
Target population (demographics)	The demographic information related to the dataset used to train, test, and validate the machine learning model.
Intervention goal	The purpose of the technological intervention such as chronic-illness treatment, management and prediction.
Chronic illness	Name of the chronic disease.
Machine learning technique	Name of the machine learning technique (e.g., random forest) used to deliver the intervention.
XAI technique	Name of the tool or algorithm (e.g., SHAP) used to explain the black-box machine learning model.
Summary of findings	A summary of the primary results obtained from the study.
Limitations	Limitations that were identified in the methodology of the study.

XAI: explainable artificial intelligence; SHAP: SHapley Additive exPlanation.

1. Study design (e.g., user study, observational study, cross-sectional study)
2. Study type:
 - QL: Qualitative
 - QN: Quantitative
 - M: Mixed methods
3. Comparative:
 - Yes: The study included a comparison group or comparative analysis
 - No: The study did not include a comparison group or comparative analysis

These classifications were visualized using a stacked bar chart to provide a clear overview of the distribution of studies across these categories. Our studies were illustrated with figures or tabulated based on target behavior (e.g., exercise, healthy eating), chronic disease (e.g., Diabetes), ML models or algorithms, XAI (Software), summary of findings, and study limitations.

Reporting bias assessment

The RoB was estimated independently by the first two authors. The final agreed-upon classification for each study

Table 3. A classification framework for assessing the risk of bias (RoB).^{27,28}

Bias	Description	Relevant domain
Selection	Differences that occur between the baseline attributes of users and groups which are compared.	Sequence generation or allocation concealment
Performance	Differences that occur in the care that is provided to each group (e.g., exposure to unrelated factors other than the intervention).	Blinding (participants and/or study personnel)
Detection	Differences that occur in how outcomes are determined between groups.	Blinding (outcome assessment)
Attrition	Difference in participant withdrawals between groups.	Incomplete or obfuscated outcome data
Reporting	Differences in reported and unreported outcomes.	Selecting reporting of outcomes (e.g., statistical analysis)
Other	Other biases that may be related to a study's design.	Contamination or carry-over effects

was reported by using Cochrane's assessment tool.²⁷ Cochrane's assessment tool was developed by an international organization consisting of 15,000 domain experts (e.g., review authors, editors, and healthcare professionals) from across 100 countries. It comprehensively outlines how to critically assess five types of biases from studies (e.g., clinical trials or randomized control trials (RCTs)) to determine the validity of outcomes. The six types of biases (selection, performance, detection, attrition, reporting, other) are summarized in Table 3.

Certainty assessment

The certainty of the evidence was assessed among 44 observational studies, 8 computation-based studies, 4 quasi-experimental studies and 2 design science research (DSR) based studies using the GRADE approach.²⁹ The final determination was made by assessing (1) the overall RoB, (2) inconsistency of results, (3) indirectness of evidence, (4) imprecision, and (5) type of study design. First, RoB based

on the five types of bias, was assessed independently and collaboratively using Cochrane's assessment tool.²⁷ Second, inconsistency of results refers to the heterogeneity in meta-analysis between all of the studies. Third, indirectness of evidence refers to whether the studies answer the research questions outlined in the article and if the findings are generalizable to the broad public. Fourth, imprecision refers to whether a study recruited enough participants which would cause a high confidence interval with relation to the estimate of effect. Lastly, the type of study design refers to whether a control group was used which determined the baseline classification of certainty of evidence. A summary of findings (SoF) table³⁰ is not included in the certainty assessment as included studies did not use a RCT design. Additionally, the differing types of study methodologies leveraged widely varying outcome measures and participants/data that do not follow a standardized measurement criterion for carrying out reasonable comparisons.

Results

The results of the systematic review are presented and described in this section using comprehensive tables, charts, and figures that address the research questions. They include key extracted characteristics for each article, publication trend, chronic disease, ML (Models/Algorithms), XAI (Software), SOFs, target behaviors, RoB, and certainty of evidence.

Study selection

The search strategy used for screening and selecting relevant articles for data extraction was based on the Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA) flowchart illustrated in Figure 1.³¹ The PRISMA 2020 checklist is available and included in the systematic review (See S1 Table in the Supplemental File). 254 articles were initially retrieved from the eight databases by using our search string. After removing 131 duplicates and 12 non-conference or journal articles, we arrived at 111 unique articles. Upon completing the Title, Abstract, and Keyword screening, 38 articles were excluded from full-text review as their subject matter was found to be unrelated to the research questions. The interrater reliability score between the first and second author was 0.7 indicating substantial agreement.³² Lastly, upon full-text review, 15 articles were excluded to arrive at 58 articles for the systematic review and final analysis. The interrater reliability score between the first and second author was 1.0 indicating an almost perfect agreement.³²

Study characteristics and results of individual studies

The study characteristics are organized into 4 groups: (1) Observational-based studies ($n=44$), (2) computation-based studies ($n=8$), (3) DSR-based studies ($n=1$), and

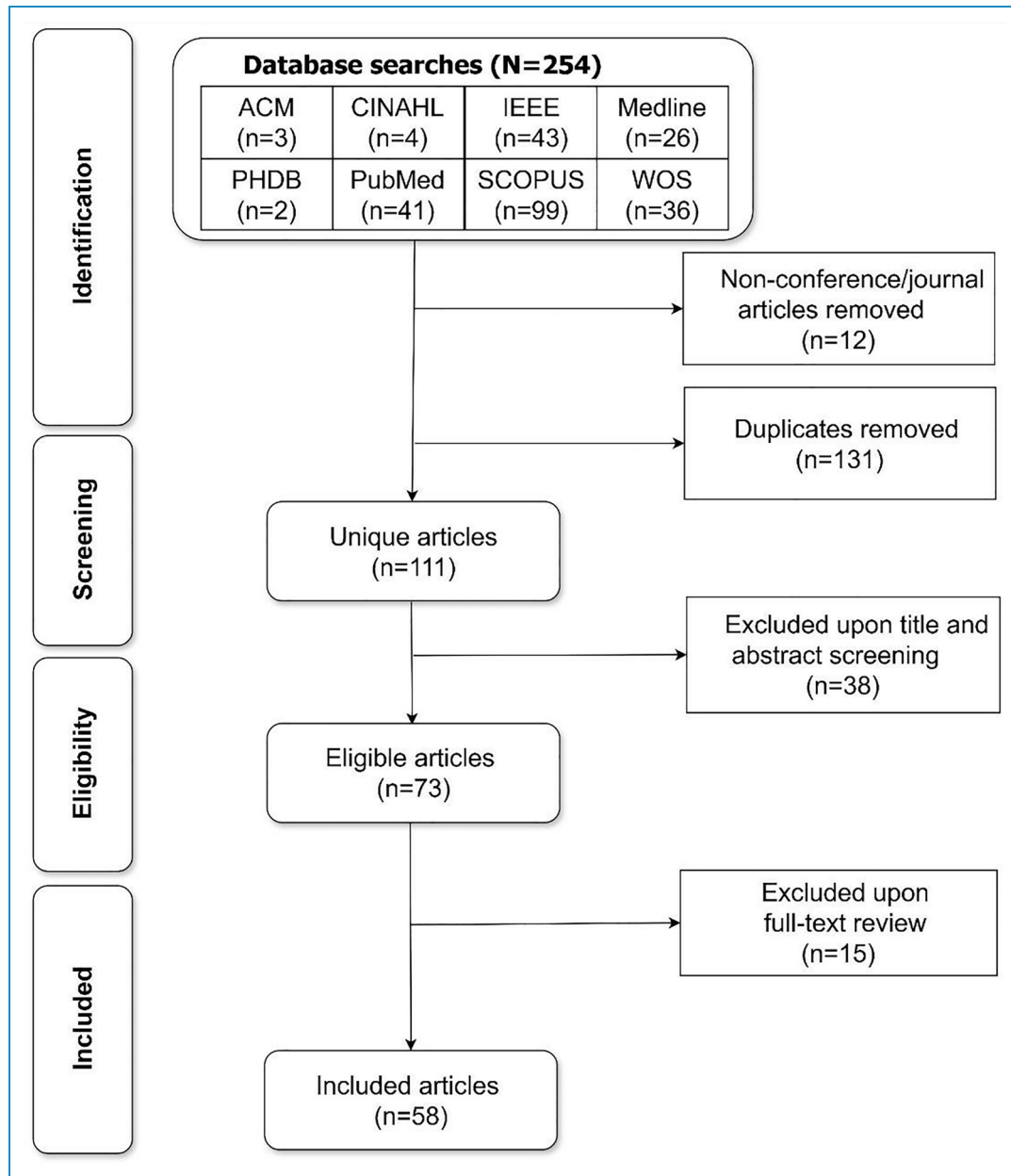


Figure 1. Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA) flowchart resulting from the screening and inclusion of articles in the systematic review.

(4) Quasi-experimental studies ($n = 5$). All of the studies in our systematic review are quantitative.

The key characteristics of each type of study are included in the systematic review (See S2 Table, S3 Table, S4 Table, and S5 Table in the Supplemental File). Each of the four tables includes information related to author identification, target population (Demographics), target behavior, chronic diseases, ML(Models/Algorithms), XAI software, study year, study type, and SoFs. and lastly, limitations.

Results of syntheses

Publication year trend. A 6-year-period ranging from 2019 to the present (2024) was identified as the active period in which our research was undertaken and findings were published. Figure 2 illustrates the publication trend, showing that the majority of the publications occurred from 2022 onwards (51/58, 87.93%). Overall, there is an upward trend in terms of the number of publications per Year from 2019 to 2023, with the only exception from 2023 to 2024 which

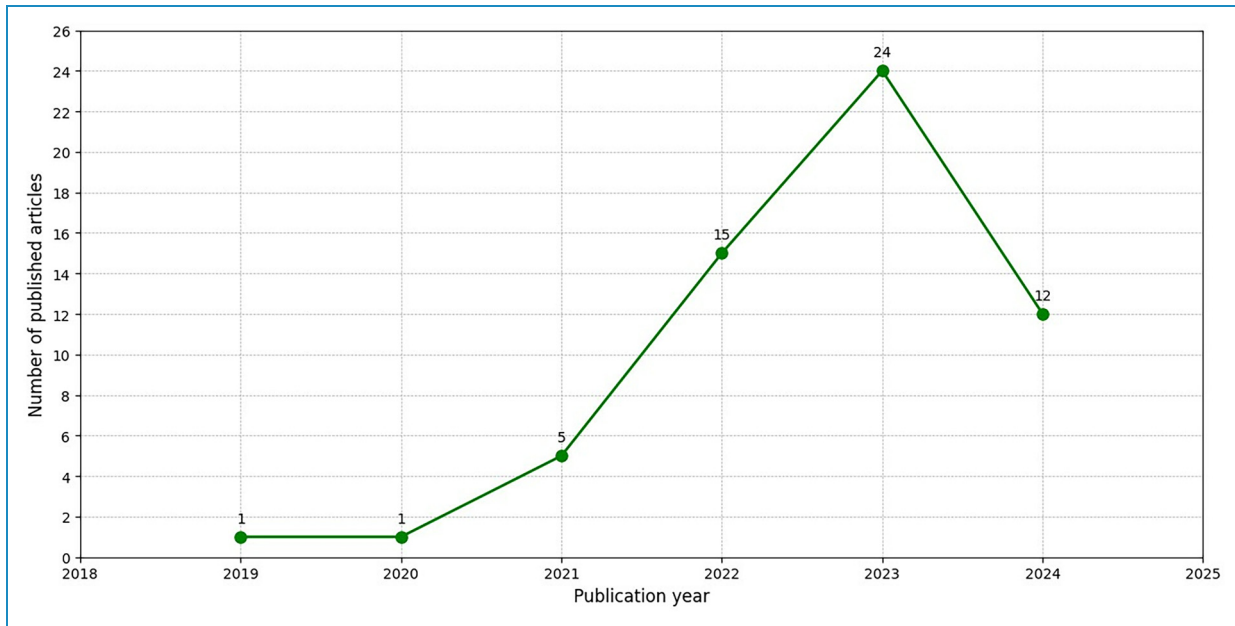


Figure 2. Distribution of published articles by year.

seems fair as our knowledge cut-off is May 2024. This overall trend is consistent with the significant increase in machine learning research within healthcare, emphasizing its evolving role in diagnostics and understanding health determinants.^{33,34}

Study location. Figure 3 illustrates the number of published articles in terms of the country of the first author. The largest contribution of work came from India (14/58, 24.14%), followed by China (5/58, 8.62%). United States and Bangladesh are each in third place (4/58, 6.90%). Finland, Italy, and the Republic of Korea share the fourth position, each contributing 3 articles (5.17% each). Several countries, including Algeria, Australia, Mauritius, Türkiye, Canada, and Spain, have 2 publications each (3.45% per country). The remaining countries, such as Colombia, Brazil, Sweden, Singapore, Pakistan, Greece, United Arab Emirates, Japan, Norway, and Germany, each contributed 1 article (1.72% each) to the total pool of publications.

Chronic diseases and corresponding XAI algorithm. Broadly speaking, a chronic condition (also known as chronic disease or chronic illness) is a health condition or disease that is persistent or otherwise long-lasting in its effects or a disease that comes with time. The term “chronic” is often applied when the course of the disease lasts for more than three months.³⁵

Table 4 outlines the different types of chronic diseases from our systematic review along with the XAI algorithm used to predict, diagnose, treat, or management of them.

The bar chart below represents the distribution of chronic diseases reported across various research papers included in our systematic review. CKD is the most frequently mentioned chronic condition, appearing in 15 studies, which accounts for 25.86% of the total. Diabetes follows with 7 mentions (12.07%), and chronic obstructive pulmonary disease (COPD) is highlighted in 5 studies (8.62%). Other diseases, such as lung disease, Alzheimer’s, and myalgic encephalomyelitis (chronic fatigue), appear less frequently, with each being mentioned in 2 or 3 studies (3.45%–5.17%).

Figure 4 shows that CKD and diabetes are the most commonly studied chronic diseases in the papers reviewed, reflecting their significant impact on public health. Respiratory conditions like COPD and general lung disease also receive considerable attention. The diversity of other chronic conditions, such as asthma, pneumonia, and obesity, is represented but with fewer mentions, suggesting varying levels of research focus across chronic diseases in the literature.

Figure 5 illustrates the frequency of usage for three XAI algorithms—LIME, SHAP, and GradCAM which are the most frequent XAI algorithms in our systematic review—across four different chronic diseases: CKD, COPD, diabetes, and lung disease. These four chronic diseases as mentioned earlier, are the most prevalent chronic diseases in our review. SHAP is the most frequently used algorithm, particularly for CKD, where it has been applied 9 times, and it also shows a significant use for COPD and diabetes, with 2 and 4 instances respectively. LIME shows a balanced usage, being applied 4 times each for CKD and diabetes, and once for lung disease. GradCAM, while used less

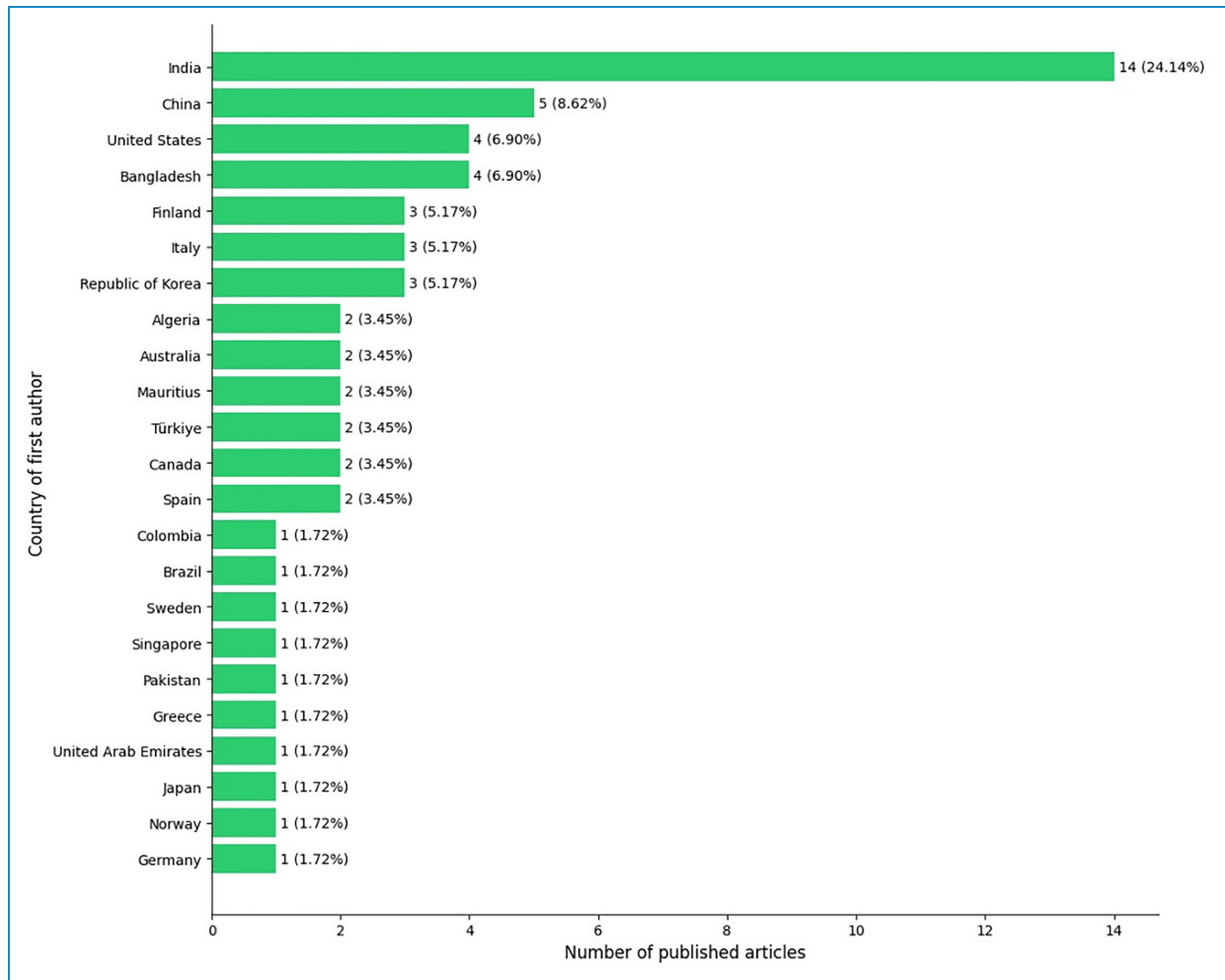


Figure 3. Distribution of published articles by the country of the first author.

overall, still has some presence, being applied twice for lung disease and COPD, and once for CKD. The chart highlights that SHAP is the dominant XAI method, particularly in the context of CKD, suggesting a strong preference or effectiveness of SHAP for explaining models related to this disease. It is evident that LIME and GradCAM have a more moderate and diversified application across the diseases.

Chronic diseases and corresponding ML algorithms. Table 5 outlines the different types of chronic diseases from our systematic review along with the ML algorithm used to predict, diagnose, treat, or management of them.

Figure 6 highlights the frequency of usage of different XAI algorithms—LIME, SHAP, and GradCAM—for predicting and diagnosing, treating, and managing the most prevalent chronic diseases in our systematic review including CKD, COPD, diabetes, and lung disease. SHAP is the most frequently utilized algorithm, particularly for CKD,

with 9 instances. LIME shows balanced usage across CKD and diabetes (both at 4 instances), while it is minimally applied to lung disease. GradCAM is less widely used but exhibits balanced usage for COPD and lung disease (both at 2 instances). Overall, SHAP demonstrates a dominant preference in CKD analysis, while other algorithms exhibit more specific but limited applications.

Chronic diseases and XAI intervention goal. Table 6 outlines the different types of chronic diseases from our systematic review along with their target behavior.

Figure 7 reveals a significant focus on prediction within the studies included in our systematic review, with 50 papers (86.21%) targeting predictive analysis of chronic diseases. This dominant interest suggests that the primary application of machine learning and XAI in this field is to anticipate disease onset or progression, emphasizing the value of early detection and proactive healthcare. In contrast, only 6 papers (10.34%)

Table 4. Chronic diseases with their corresponding XAI algorithm and authors.

Chronic disease	XAI algorithm	Author(s)
Alzheimers	PDA PGM	Kamal MS et al. ³⁶
Alzheimers	SHAP and LIME	Sekaran K et al. ³⁷
Anemia(s)	XGBoost, logistic regression, decision tree, Naive Bayes, SVM, KNN, random forest, explainable boosting machine	J. Prajapati et al. ³⁸
Asthma	Logic learning machine (rule-based)	Narteni S et al. ³⁹
Breast cancer	LIME	I. Maouche et al. ⁴⁰
Chronic kidney disease	TreeExplainer (Shapley value explanations)	Lundberg SM et al. ⁴¹
Chronic kidney disease	SHAP	K.Jhumka et al. ⁴²
Chronic kidney disease	SHAPs	VN Manju et al. ⁴³
Chronic kidney disease	PFI, PDP, SHAPs	P. A. Moreno-Sanchez ²⁶
Chronic kidney disease	LIME	A. Vijayvargiya et al. ⁴⁴
Chronic kidney disease	LIME SHAPs SKATER Metrics: Interpretability, fidelity, fidelity-to-interpretability ratio	M. A. Islam et al. ⁴⁵
Chronic kidney disease	Grad-CAM	M. H. K. Mehedi et al. ⁴⁶
Chronic kidney disease	SHAPs	K. Jhumka et al. ⁴⁷
Chronic kidney disease	Graph properties	Y. Shang et al. ⁴⁸
Chronic kidney disease	SHAPs	S. Paul et al. ⁴⁹
Chronic kidney disease	Case-based reasoning model	G. R. Vásquez-Morales ⁵⁰
Chronic kidney disease	SHAPs LIME	Ghosh SK, Khandoker AH ⁵¹
Chronic kidney disease	SHAPs	Hamed O et al. ⁵²
Chronic kidney disease	LIME	Arumugham V et al. ⁵³
Chronic kidney disease	Metrics: Interpretability, fidelity, fidelity-to-interpretable ratio Interpretability: Input features that the model presents as the explanation (via masking). If the model achieves a correct classification after masking some portion of the input features, the remaining features have a significant semantic value. Fidelity: Relation of classification metric results of the full-interpretable model and its uninterpretable counterpart (baseline). Fidelity shows the percentage of initial performance retained by	P. A. Moreno-Sanchez ⁵⁴

(continued)

Table 4. Continued.

Chronic disease	XAI algorithm	Author(s)
	becoming interpretable. Fidelity-to-interpretable ratio: Refers to how much of the model's interpretability is sacrificed for performance. We aim for 0.5 as a good metric.	
Chronic obstructive pulmonary disease	Grad-CAM SHAPs	A. V. Ikechukwu; S. Murali ⁵⁵
Chronic obstructive pulmonary disease	LRP Grad-CAM	A. V. Ikechukwu; M. S; H. B ⁵⁶
Chronic obstructive pulmonary disease	SHAPs PDP	Wang X et al. ⁵⁷
Chronic obstructive pulmonary disease	LLM	Vaccari I et al. ⁵⁸
Chronic obstructive pulmonary disease	Case-based reasoning model	El-Magd L.M.A et al. ⁵⁹
Diabetes	PUTreeLIME In contrast to existing interpretation methods, PUTreeLIME takes into account the interaction among different community models within our PUTree. Unlike LIME, we employ the uncertainty sampling method to sample instances which ensures that the sampled instances have a higher degree of responsibility for the model being explained.	Wu et al. ⁶⁰
Diabetes	Most-weighted path, Most-weighted combination, Maximum frequency difference explanation	A. Ahmad; S. Mehruz ⁶¹
Diabetes	SHAPs	I. Shaheen et al. ⁶²
Diabetes	LIME	P Nagaraj et al. ⁶³
Diabetes	SHAPs LIME	Kibria HB ⁶⁴
Diabetes	Attention Mechanism SHAPs LIME	Joseph LP et al. ⁶⁵
Diabetes	SHAPs LIME	Ganguly R, Singh D. ⁶⁶
Fibromyalgia	PDP MDI	Moreno-Sanchez et al. ⁶⁷
General	SHAPs LIME	Huang M et al. ⁶⁸
General (heart disease/cardiovascular disease/ lung disease)	Additive explanatory factors (EBM-produced)	Pfeuffer N. ⁶⁹
Hematological malignant disorder	SHAPs	Rodríguez-Belenguer P. ⁷⁰

(continued)

Table 4. Continued.

Chronic disease	XAI algorithm	Author(s)
Hepatic steatosis non-alcoholic fatty liver disease	Partial dependency	R. Deo; S. Panigrahi ⁷¹
Knee osteoarthritis	Grad-CAM	R. Ahmed; A. S. Imran ⁷²
Leprosy	Visualization of activation layer, occlusion sensitivity, Grad-CAM	A. K. Baweja et al. ⁷³
Liver cirrhosis	SHAPs	G. Arya et al. ⁷⁴
Lung disease	Grad-CAM	R. Sharma et al. ⁷⁵
Lung disease	LIME	S. P. Koyyada et al. ⁷⁶
Lung disease	Attention Grad-CAM	Choi Y et al. ⁷⁷
Lymphocytic leukemia	SHAPs	Morabito F et al. ⁷⁸
Lymphomagenesis	SHAPs	V. C. Pezoulas et al. ⁷⁹
Lymphomas	Anchor: Generates explanations for specific diagnoses. SHAP: Ranks descriptors according to their global influence on model decisions.	T. P. De Faria ⁸⁰
Metabolic syndrome	Explanation module	Benmohammed K et al. ⁸¹
Myalgic encephalomyelitis (chronic fatigue)	SHAPs TreeMap	Yagin FH et al. ⁸²
Myalgic encephalomyelitis (chronic fatigue)	SHAPs	Yagin FH et al. ⁸³
Non-communicable diseases	DeepSHAP	K. Davagdorj et al. ⁸⁴
Obesity	Graph properties, SHAP	Brakefield WS et al. ⁸⁵
Obstructive sleep apnea	LIME	Troncoso-García A.R et al. ⁸⁶
Parkinsons	SHAPs LIME SHAPASH	Junaid M et al. ⁸⁷
Pneumonia	Attention mechanism	Ukwuoma CC et al. ⁸⁸
Pulmonary (normal, COVID-19, tuberculosis, pneumonia)	LRP	G. Marvin et al. ⁸⁹
Ulcers	Grad-CAM, LIME, and SHAP were tested as viable explainability methods, with SHAP used as the preferable explainability method.	Lo et al. ⁹⁰
Usual interstitial pneumonia	The physician who labels the clusters serves as the XAI component of the non-end-to-end system.	Uegami et al. ⁹¹
Asthma, bronchiectasis, bronchiolitis, chronic	Fuzzy ID3	J. Li et al. ⁹²

(continued)

Table 4. Continued.

Chronic disease	XAI algorithm	Author(s)
obstructive pulmonary disease, lower respiratory tract infection, upper respiratory tract infection, pneumonia (a set of respiratory illnesses)		

XAI: explainable artificial intelligence; PDA: pixel density analysis; PGM: probabilistic graphical model; SHAP: SHapley Additive exPlanation; LIME: local interpretable model-agnostic explanation; XGBoost: extreme gradient boosting; SVM: support vector machine; KNN: k-nearest neighbor; PFI: permutation feature importance; PDP: partial dependence plot; Grad-CAM: gradient-weighted class activation mapping; LRP: layer-wise relevance propagation; LLM: logic learning machine; PUTreeLIME: positive-unlabeled tree LIMEs; MDI: mean decrease impurity; DeepSHAP: deep SHapley Additive exPlanation.

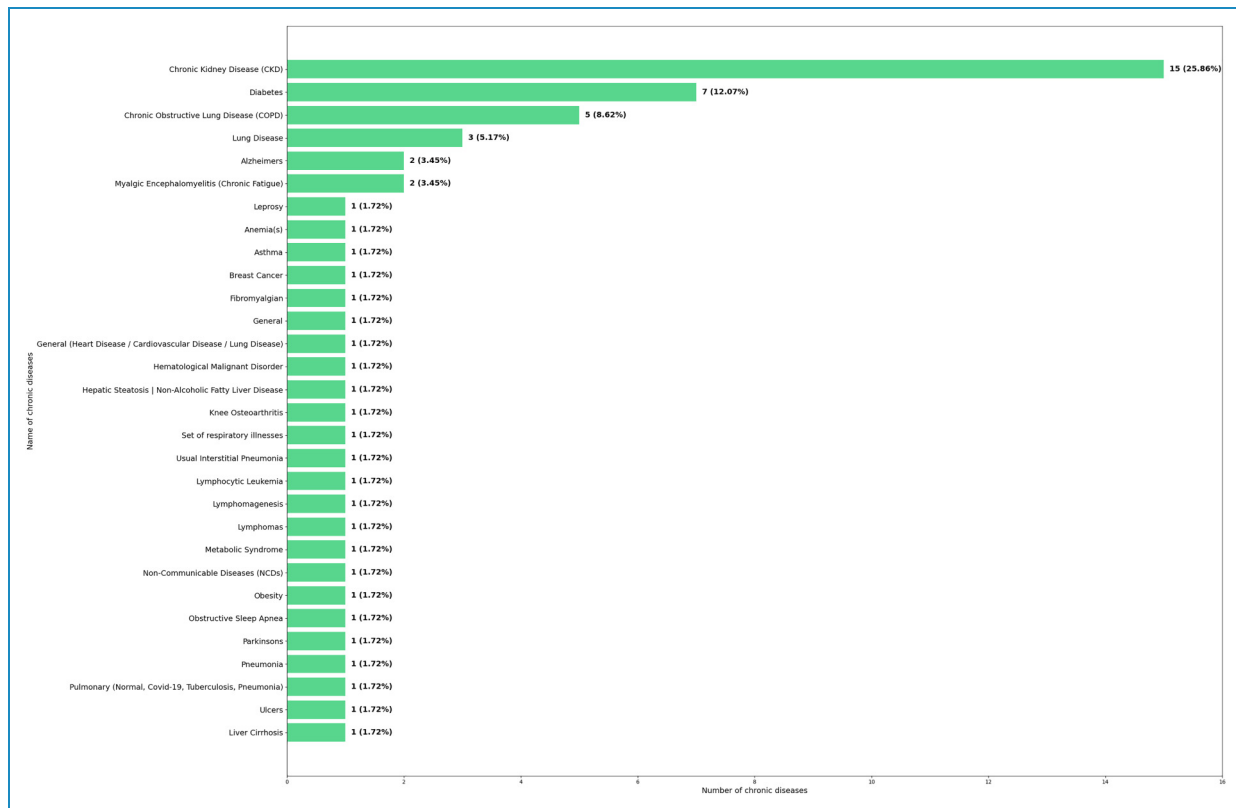


Figure 4. Prevalence of chronic diseases in our systematic review.

concentrate on disease management, indicating a smaller yet notable interest in improving patient care processes or monitoring. The least studied aspect is treatment, with just 2 papers (3.45%) focusing on treatment-related outcomes, highlighting a potential area for future research to explore how machine learning and AI can contribute directly to therapeutic interventions.

Figure 8 displays the percentage usage of different XAI technological goals—prediction, management, and treatment—applied to the four most prevalent chronic diseases in our systematic review. The findings reveal a strong

focus on prediction across all diseases, especially for diabetes and lung disease, where prediction is applied 100% of the time. CKD also shows a high percentage of prediction (93.3%), with a small portion allocated to treatment (6.7%). Similarly, for COPD, 80% of cases focus on prediction, while 20% are dedicated to management. Notably, none of the diseases show a substantial emphasis on treatment, suggesting that research predominantly focuses on predictive approaches rather than treatment interventions for these chronic conditions. This trend underscores the primary role of machine learning in forecasting disease progression or onset, rather than direct

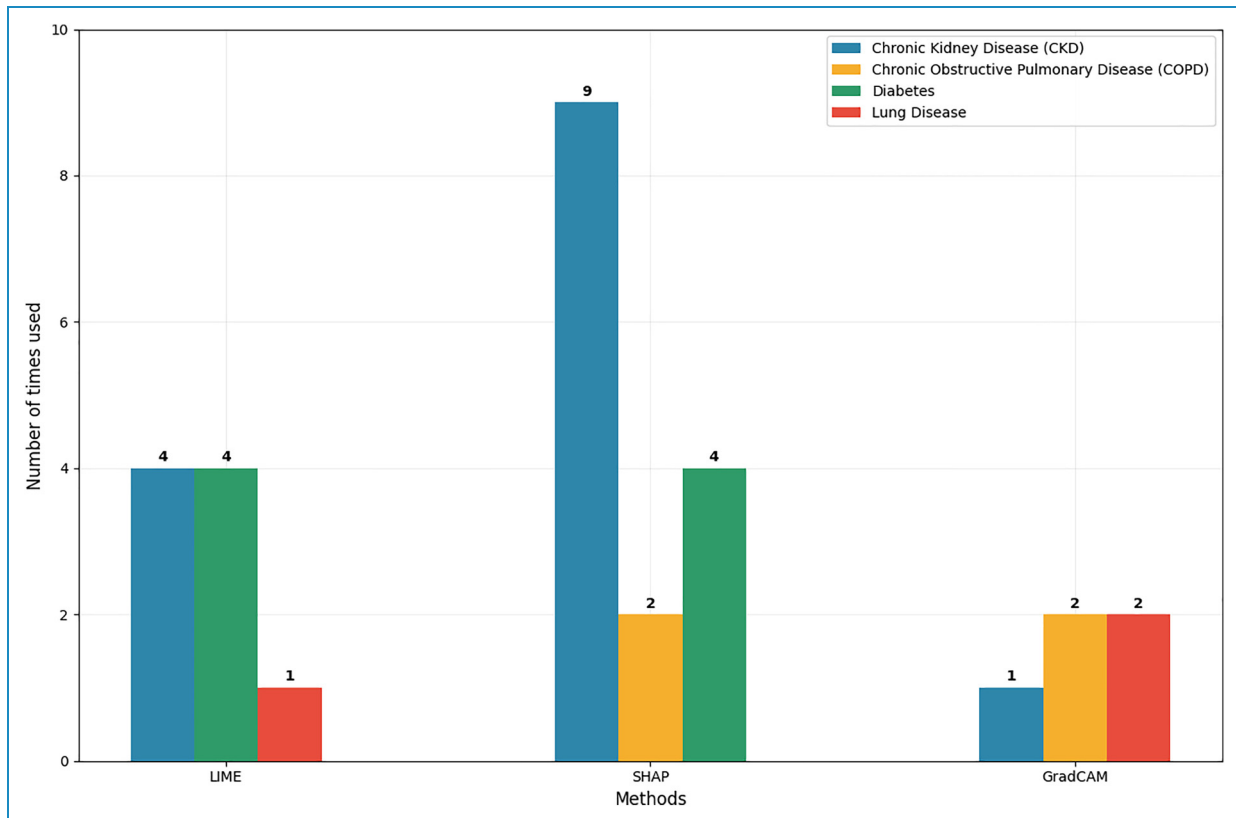


Figure 5. Frequency of the most prevalent explainable artificial intelligence (XAI) algorithms used in predicting, diagnosing, treating, and managing chronic diseases.

management or therapeutic applications in the reviewed studies.

RoB in studies. Figures 9, 10, 11A, 11B, 11C, 12, and 13 show the traffic-light plot of the risks of bias for the three study types: Computational, observational, quasi-experimental, and DSR.

Overall, as seen in Figure 13, except for one paper, all the included studies were low in RoB.

Certainty of evidence. The certainty of evidence assessment for the included studies revealed varying levels of methodological rigor and reliability across the reviewed literature. Among the 58 papers analyzed, 8 studies (13.8%) demonstrated high certainty of evidence, characterized by robust methodological approaches, comprehensive clinical validations, and thorough testing with diverse patient populations in real healthcare settings. The majority of studies (48 papers, 82.8%) showed moderate certainty of evidence, suggesting generally sound methodologies with some limitations. Only 2 studies (3.4%) were assessed as having low certainty of evidence, as evidenced by specific methodological limitations. These included Pfeuffer et al. (2021),⁶⁹ which based findings on just 43 usable samples from a single patient over three months and

validation from only three general practitioners. The second study, Vaccari et al. (2021),⁵⁸ raised particular concerns in its evaluation methodology, employing metrics (Jensen-Shannon divergence and Fréchet Inception Distance) originally designed for image generation without proper validation for medical data. The authors acknowledged this limitation, noting these metrics were borrowed from image processing applications as “a starting point.” This approach contrasts sharply with high-certainty medical studies that employ clinically validated, domain-specific metrics ensuring preservation of diagnostic features and statistical equivalence of clinical parameters. The FID metric’s assumption of Gaussian distributions, while suitable for image analysis, remains unvalidated for medical time series data from IoMT devices, where subtle physiological variations can have significant clinical implications. When compared to established research standards, where high-quality healthcare AI studies typically involve hundreds or thousands of patients across multiple clinical sites with comprehensive validation phases and domain-appropriate evaluation metrics, these two studies fell significantly short. Despite these limitations in a small portion of the reviewed literature, the predominance of moderate and high-certainty studies suggests a generally robust body of evidence in the field, though continued emphasis on rigorous methodology

Table 5. Chronic diseases with their corresponding ML algorithm and authors.

Chronic disease	ML algorithm	Author(s)
Alzheimers	SNN	Kamal MS et al. ³⁶
Alzheimers	LR, RF, L-SVMs, NB, MLP-NN	Sekaran K et al. ³⁷
Anemia(s)	XGBoost, LR, decision tree, NB, SVM, KNN, RF, EBM	J. Prajapati et al. ³⁸
Asthma	Logic learning machine	Narteni S et al. ³⁹
Breast cancer	CatBoost	I. Maouche et al. ⁴⁰
Chronic kidney disease	Gradient boosted decision trees	Lundberg SM et al. ⁴¹
Chronic kidney disease	XGBoost, AdaBoost, DNN	Jhumka et al. ⁴²
Chronic kidney disease	Tree-based explainable AI (DT-EAI) (decision tree)	VN Manju et al. ⁴³
Chronic kidney disease	Decision trees: RF, Extra Trees, AdaBoost, XGBoost	P. A. Moreno-Sánchez ⁵⁴
Chronic kidney disease	RF, SVM, decision tree, LR, KNN	A. Vijayvargiya et al. ⁴⁴
Chronic kidney disease	Decision tree, RF	M. A. Islam et al. ⁴⁵
Chronic kidney disease	CNN, DNN VGG16, MobileNetV2, InceptionV3	M. H. K. Mehedi et al. ⁴⁶
Chronic kidney disease	XResNet50 (CNN)	K. Jhumka et al. ⁴⁷
Chronic kidney disease	Knowledge graphs (layered modules): OMOP CDM is a standard data schema for normalizing heterogeneous EHR datasets to achieve multicenter collaborative research. The system uses EHR data that underwent the extraction, transformation, and loading (ETL) process into the OMOP CDM format.	Y. Shang et al. ⁴⁸
Chronic kidney disease	KNN, NB, SVM	S. Paul et al. ⁴⁹
Chronic kidney disease	Neural Networks	G. R. Vázquez-Morales et al. ⁵⁰
Chronic kidney disease	LR, RF, Decision Tree, NB, XGBoost	Ghosh SK, Khandoker AH ⁵¹
Chronic kidney disease	LSTM	Hamed O et al. ⁵²
Chronic kidney disease	DNN	Arumugham V et al. ⁵³
Chronic kidney disease	Decision Tree, RF, Extra Trees, Adaboost, Gradient Boosting, XGBoost, Ensemble Voting Classifier (Hard VotingClassifier)	P. A. Moreno-Sanchez ²⁶
Chronic obstructive pulmonary disease	Xception ResNet50V2 CNNs (transfer learning)	A. V. Ikechukwu; S. Murali ⁵⁵

(continued)

Table 5. Continued.

Chronic disease	ML algorithm	Author(s)
Chronic obstructive pulmonary disease	Chronic obstructive pulmonary diseaseNet (ResNet 50) Finetuned (+ Transfer Learning with ImageNet weights) ResNet50 (Base)	A. V. Ikechukwu; M. S; H. B ⁵⁶
Chronic obstructive pulmonary disease	CatBoost, NGBoost, XGBoost, LightGBM, Random Forest, SVM, LR	Wang X et al. ⁵⁷
Chronic obstructive pulmonary disease	GAN decision tree	Vaccari I et al. ⁵⁸
Chronic obstructive pulmonary disease	ResNet18, ResNet34, ResNet50, AlexNet, GoogleNet	El-Magd L.M.A et al. ⁵⁹
Diabetes	PUTree	Wu et al. ⁶⁰
Diabetes	MLP, weighted KNN, SVM	A. Ahmad; S. Mehruz ⁶¹
Diabetes	Hybrid #1: Highway + LeNet = Hi-Le + LeNet Hybrid #2: Highway + LeNet + TCN = HiTCLe	I. Shaheen et al. ⁶²
Diabetes	SVM, RF, decision tree, XGBoost	P Nagaraj et al. ⁶³
Diabetes	ANN, RF, SVM, LR, AdaBoost, XGBoost	Kibria HB ⁶⁴
Diabetes	TabNet	Joseph LP et al. ⁶⁵
Diabetes	Ensemble (unknown)	Ganguly R.; Singh D. ⁶⁶
Fibromyalgia	Decision trees, LR, SVM, RF, AdaBoost, Extra Trees, RUBOOST	Moreno-Sanchez PA et al. ⁶⁷
General	Singular value decomposition, decision tree, GCNs, XGBoost	Huang M et al. ⁶⁸
General (heart disease/cardiovascular disease/ lung disease)	Binary EBM	Pfeuffer N. ⁶⁹
Hematological malignant disorder	LR, decision tree, RF, SVM	Rodríguez-Belenguer P. ⁷⁰
Hepatic steatosis non-alcoholic fatty liver disease	SVM	R. Deo; S. Panigrahi ⁷¹
Knee osteoarthritis	CNN (fine-tuned, pre-trained): VGG16, VGG19, ResNet-50, ResNet-101, EfficientNetb7	R. Ahmed; A. S. Imran ⁷²
Leprosy	AXI-CNN (LeprosyNet)	A. K. Baweja et al. ⁷³
Liver cirrhosis	XGBoost, LR	G. Arya et al. ⁷⁴
Lung disease	GANs + CNNs (VGG-16, VC-Net architectures)	R. Sharma et al. ⁷⁵
Lung disease	CNN	S. P. Koyyada et al. ⁷⁶
Lung disease	ECA-Net (modified) VGGish LACM	Choi Y et al. ⁷⁷

(continued)

Table 5. Continued.

Chronic disease	ML algorithm	Author(s)
Lymphocytic leukemia	NN	Morabito F et al. ⁷⁸
Lymphomagenesis	Multinomial NB (FMNB), multilayer perceptron (FMLP), SVM (FSCVM), gradient boosting trees (FGBT, FDART)	V. C. Pezoulas et al. ⁷⁹
Lymphomas	Linear regression, SVM, GBDT	T. P. De Faria ⁸⁰
Metabolic syndrome	ANN, SVM, KNN, NB	Benmohammed K et al. ⁸¹
Myalgic encephalomyelitis (chronic fatigue)	RF, Gaussian NB, gradient boosting classifier, LR, RF classifier	Yagin FH et al. ⁸³
Myalgic encephalomyelitis (chronic fatigue)	XGBoost, SVC, LR, RF, decision tree, NB	Yagin FH et al. ⁸²
NCDs	DNN	K. Davagdorj et al. ⁸⁴
Obesity	SVR machine model Knowledge graph	Brakefield WS et al. ⁸⁵
Obstructive sleep apnea	LR, K-nearest neighbors, decision tree, RF, gradient boosting classifier	Troncoco-García A.R et al. ⁸⁶
Parkinsons	SVM, RF, ETC, LGBM, SGD	Junaid M et al. ⁸⁷
Pneumonia	Ensemble A: DenseNet201, VGG16, GoogleNet Ensemble B: DenseNet201, InceptionResNetV2, Xception Transformer encoder (with self-attention mechanism)	Ukwuoma CC et al. ⁸⁸
Pulmonary (normal, COVID-19, tuberculosis, pneumonia)	Deep CNN + Base/Transfer Cases a.k.a. Deep & transfer learning	G. Marvin et al. ⁸⁹
Ulcers	(1) Models used during the development process include different versions of DenseNet, MobileNet, and ResNet for classification (2) DeepLab, FPN, or U-Net for segmentation	Lo et al. ⁹⁰
Usual interstitial pneumonia	Feature extractor (Tile classification & visualization): ResNet18 CNN + Transfer Learning UIP Prediction by MIXTURE: RF, SVM	Uegami et al. ⁹¹
Asthma, bronchiectasis, bronchiolitis, Chronic obstructive pulmonary disease, lower respiratory tract infection, upper respiratory tract infection, pneumonia (a set of respiratory illnesses)	Ensemble knowledge distillation with Fuzzy logic (multiple teacher–student models) Teacher → Classification Network Student → Network CNN's + Fuzzy decision tree	J. Li et al. ⁹²

SNN: spike neural network; LR: logistic regression; RF: random forest; L-SVM: linear support vector machine; NB: Naive Bayes; MLP-NN: multilayered perceptron neural network; DNN: deep neural network; DT-EAI: decision tree-based EAI; AdaBoost: adaptive boosting; XGBoost: extreme gradient boosting; ANN: artificial neural network; SVM: support vector machine; GCN: graph convolutional network; EBM: explainable boosting machine; ECA: efficient channel attention; LACM: light attention connected module; NN: neural network; GBDT: gradient boosting decision trees; NCDs: non-communicable diseases; SVR: support vector regression; ETC: Extra Trees Classifier; LGBM: light gradient boosting machines; SGD: stochastic gradient descent; CNN: convolutional neural network; LSTM: long short-term memory; GAN: generative adversarial network; PUTree: positive-unlabeled learning tree; MLP: multilayer-perceptron.

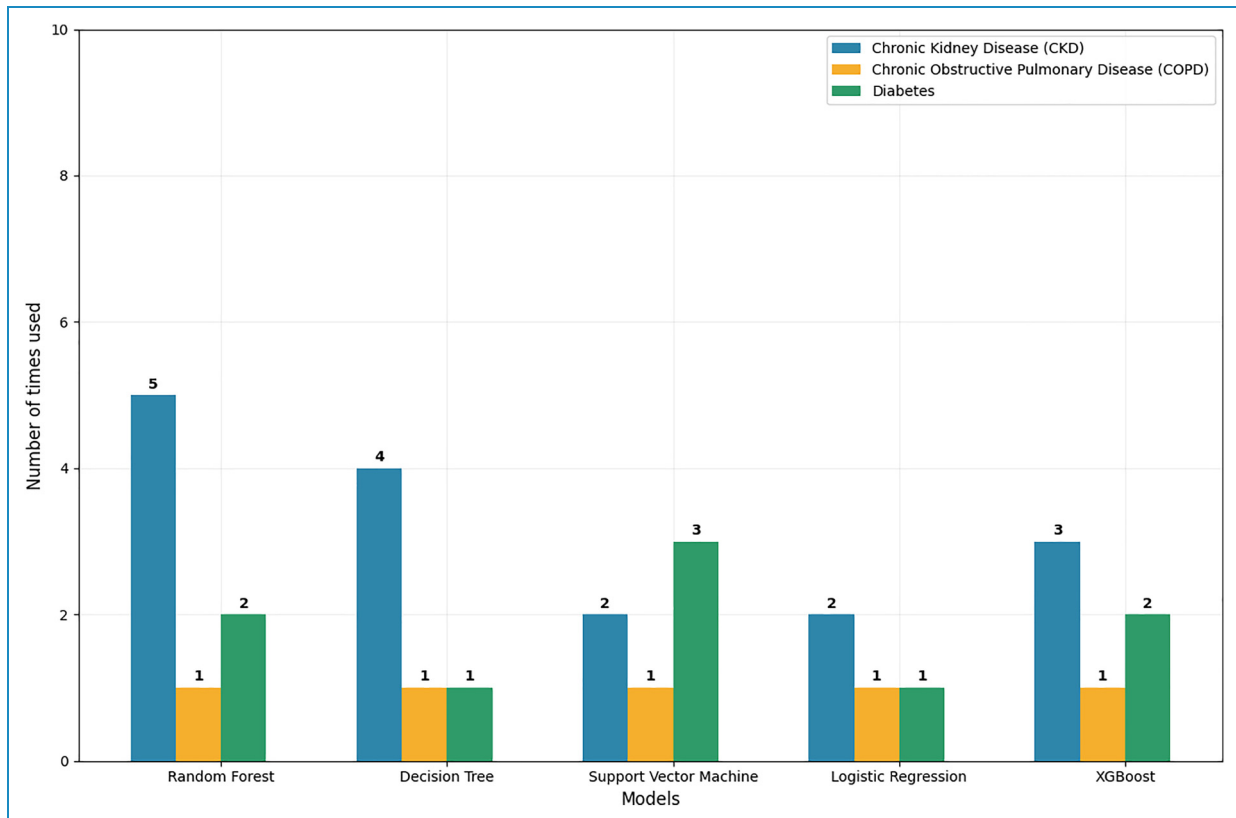


Figure 6. Frequency of the most prevalent ML algorithms used in predicting, diagnosing, treating, and managing chronic diseases.

and comprehensive clinical validation would further strengthen the evidence base.

Discussion

The prediction category encompasses studies focused on early detection and risk assessment, representing the majority of current research efforts (50/58, 86.2%). The treatment category is notably smaller, primarily represented by medication recommendation systems and resource utilization prediction for treatment planning, with limited exploration of direct therapeutic intervention optimization. Management includes papers addressing ongoing monitoring and clinical decision-support tools for existing conditions. This imbalanced distribution reflects the current state of XAI research in chronic disease care, where prediction applications significantly outweigh management and treatment applications. By examining these distinct yet interconnected domains separately, we can better understand the nuanced applications of XAI across different contexts within chronic disease care. Moreover, CKD and Diabetes emerged as the most frequently studied conditions in our systematic review, a trend that is clearly illustrated in Figure 4, which shows the prevalence of chronic diseases across the included studies.

XAI algorithms and the prediction of chronic diseases

Our systematic review revealed that three XAI models—SHAP,¹³ LIME,¹² and Grad-CAM⁹⁴—dominate the landscape of XAI applications in chronic disease care. LIME and SHAP share common advantages that make them particularly suitable for healthcare applications. Both methods are model-agnostic, meaning they can explain predictions from any machine learning model, which is crucial in healthcare where various modeling approaches may be employed. Grad-CAM's prominence, particularly in diagnostic applications, can be attributed to its effectiveness with convolutional neural networks (CNNs), which are widely used in medical imaging.⁹⁵ The technique's ability to generate visual explanations by highlighting regions of interest in medical images makes it especially valuable for conditions requiring radiological assessment, such as diabetic retinopathy.⁹⁶

The convergence of these three methods suggests a maturation in the field of healthcare XAI, where practitioners have identified tools that balance technical robustness with clinical utility. However, this concentration also raises important considerations. A notable limitation of all three methods is their focus on feature attribution and local explanations, potentially missing higher-level patterns or causal relationships that could be crucial in chronic disease management.

Table 6. Chronic diseases with their corresponding XAI intervention goal.

Chronic disease	XAI intervention goal	Author(s)
Alzheimers	Prediction of disease	Kamal MS et al. ³⁶ , Sekaran K et al. ³⁷
Anemias	Prediction	J. Prajapati et al. ³⁸
Asthma	Management of disease	Narteni S et al. ³⁹
Breast cancer	Prediction	I. Maouche et al. ⁴⁰
Chronic kidney disease	Prediction	Lundberg SM et al. ⁴¹
Chronic kidney disease	Prediction	Jhumka et al. ⁴²
Chronic kidney disease	Prediction	VN Manju et al. ⁴³
Chronic kidney disease	Prediction	P. A. Moreno-Sánchez ²⁶
Chronic kidney disease	Prediction	A. Vijayvargiya et al. ⁴⁴
Chronic kidney disease	Prediction	M. A. Islam et al. ⁴⁵
Chronic kidney disease	Prediction	M. H. K. Mehedi et al. ⁴⁶
Chronic kidney disease	Prediction	K. Jhumka et al. ⁴⁷
Chronic kidney disease	Prediction	Y. Shang et al. ⁴⁸
Chronic kidney disease	Prediction	S. Paul et al. ⁴⁹
Chronic kidney disease	Prediction	G. R. Vázquez-Morales et al. ⁵⁰
Chronic kidney disease	Prediction	Ghosh SK, Khandoker AH ⁵¹
Chronic kidney disease	Treatment of disease	Hamed O et al. ⁵²
Chronic kidney disease	Prediction	Arumugham V et al. ⁵³
Chronic kidney disease	Prediction	P. A. Moreno-Sanchez ²⁶
Chronic obstructive pulmonary disease	Prediction	A. V. Ikechukwu; S. Murali ⁵⁵
Chronic obstructive pulmonary disease	Prediction	A. V. Ikechukwu; M. S; H. B ⁵⁶
Chronic obstructive pulmonary disease	Prediction	Wang X et al. ⁵⁷
Chronic obstructive pulmonary disease	Management of disease	Vaccari I et al. ⁵⁸
Chronic obstructive pulmonary disease	Prediction	El-Magd L.M.A et al. ⁵⁹
Diabetes	Prediction	Wu et al. ⁶⁰
Diabetes	Prediction	A. Ahmad; S. Mehruz ⁶¹
Diabetes	Prediction	I. Shaheen et al. ⁶²
Diabetes	Prediction	P Nagaraj et al. ⁶³

(continued)

Table 6. Continued.

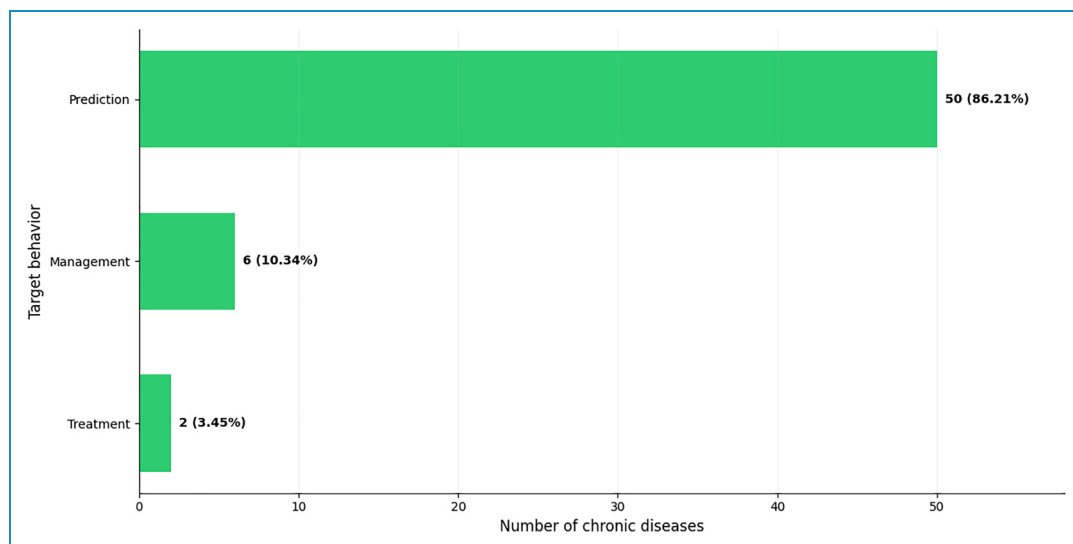
Chronic disease	XAI intervention goal	Author(s)
Diabetes	Prediction	Kibria HB ⁶⁴
Diabetes	Prediction	Joseph LP et al. ⁶⁵
Diabetes	Prediction	Ganguly R.; Singh D. ⁶⁶
Fibromyalgia	Management of disease	Moreno-Sanchez PA et al. ⁶⁷
General	Treatment of disease	Huang M et al. ⁶⁸
General (heart disease/cardiovascular disease/lung disease)	Management of disease	Pfeuffer N. ⁶⁹
Hematological malignant disorder	Prediction	Rodríguez-Belenguer P. ⁷⁰
Hepatic steatosis non-alcoholic fatty liver disease	Prediction	R. Deo; S. Panigrahi ⁷¹
Knee osteoarthritis	Prediction	R. Ahmed; A. S. Imran ⁷²
Leprosy	Prediction	A. K. Baweja et al. ⁷³
Liver cirrhosis	Prediction	G. Arya et al. ⁷⁴
Lung disease	Prediction	R. Sharma et al. ⁷⁵
Lung disease	Prediction	S. P. Koyyada et al. ⁷⁶
Lung disease	Prediction	Choi Y et al. ⁷⁷
Lymphocytic leukemia	Prediction	Morabito F et al. ⁷⁸
Lymphomagenesis	Prediction and management of disease	V. C. Pezoulas et al. ⁷⁹
Lymphomas	Prediction	T. P. De Faria ⁸⁰
Metabolic syndrome	Prediction	Benmohammed K et al. ⁸¹
Myalgic encephalomyelitis (chronic fatigue)	Prediction	Yagin FH et al. ⁸³
Myalgic encephalomyelitis (chronic fatigue)	Prediction	Yagin FH et al. ⁸²
NCDs	Prediction	K. Davagdorj et al. ⁸⁴
Obesity	Prediction and management of disease	Brakefield WS et al. ⁸⁵
Obstructive sleep apnea	Prediction	Troncoso-García A.R et al. ⁸⁶
Parkinsons	Prediction	Junaid M et al. ⁸⁷
Pneumonia	Prediction	Ukwuoma CC et al. ⁸⁸
Pulmonary (normal, Covid-19, tuberculosis, pneumonia)	Prediction	G. Marvin et al. ⁸⁹
Ulcers	Prediction	Lo et al. ⁹⁰

(continued)

Table 6. Continued.

Chronic disease	XAI intervention goal	Author(s)
Usual interstitial pneumonia	Prediction	Uegami et al. ⁹¹
Asthma, bronchiectasis, bronchiolitis, chronic obstructive pulmonary disease, lower respiratory tract infection, upper respiratory tract infection, pneumonia (a set of respiratory illnesses)	Prediction	J. Li et al. ⁹²

NCD: non-communicable disease.

**Figure 7.** XAI intervention goal of the studies in our systematic review.

In our systematic review, 50/58 papers (86.2%) are focused on the prediction of chronic diseases. In the following sections, we first go through the XAI algorithms used in the prediction of chronic diseases.

Applications of SHAP in the prediction of CKD

Jhumka et al.⁴² used SHAP to investigate feature contributions in a deep neural network (DNN) model for CKD prediction. The authors specifically employed SHAP to validate if the model's predictions aligned with medical explanations. They found that specific gravity, diabetes, and albumin were key predictors, demonstrating how SHAP can bridge the gap between machine learning predictions and clinical understanding. However, The dataset used is relatively small ($n=400$ rows) and the features ($n=24$) are not generalizable to other datasets. This makes comparison difficult against other sources of information.

Manju et al.⁴³ utilized SHAP for evaluating a decision tree-based EAI (DT-EAI) approach. SHAP was used to

calculate Shapley values for measuring feature importance in the decision-making process. The study emphasized the role of SHAP in enhancing transparency and clinical decision-making, particularly in interpreting how different attributes contribute to CKD diagnosis.

Jhumka et al.⁴⁷ incorporated SHAP to provide visual explanations of their model's predictions, using it to analyze per-class feature significance. The study demonstrated how SHAP can help identify the most relevant features that determine CKD, making the complex deep-learning model more interpretable for healthcare professionals.

Paul et al.⁴⁹ employed SHAP as a model explanation method to disclose the importance of features in CKD classification. They used multiple SHAP visualization techniques (bar plots, swarm plots, and waterfall plots) to provide comprehensive insights into how different features contribute to the model's predictions. The study showed that GFR, age, and sex had the highest impact on CKD prediction. The paper's limitations are that the dataset used is static and also class imbalance is evident.

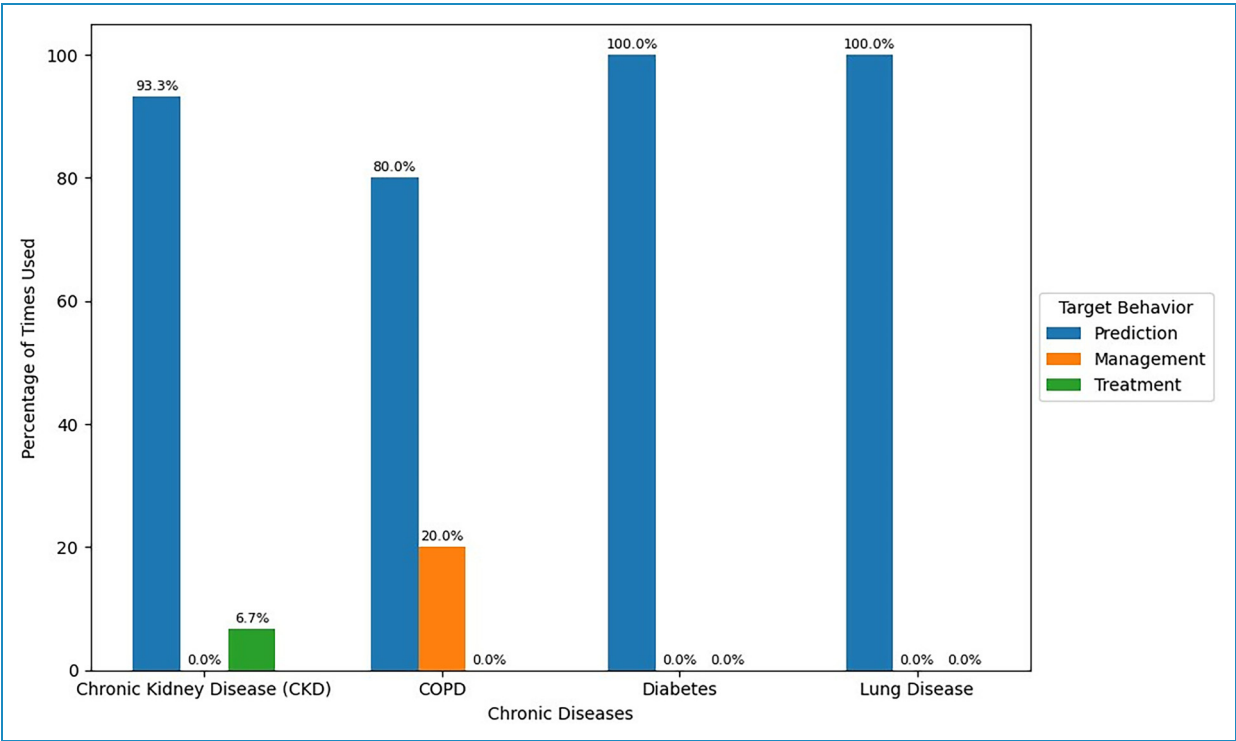


Figure 8. Distribution of prediction, management, and treatment focus across four most prevalent chronic diseases in our systematic review.



Figure 9. Traffic-light plot of the risk of bias (RoB) from computational studies.⁹³

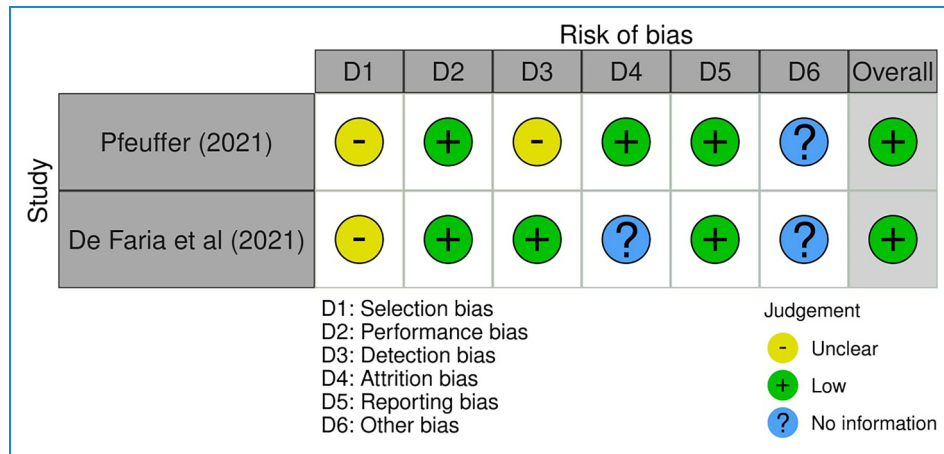


Figure 10. Traffic-light plot of the RoB from design research science studies. DSR: design science research⁹³; RoB: risk of bias.

Applications of SHAP in prediction of COPD

Ikechukwu et al.⁵⁵ utilized SHAP alongside Grad-CAM to explain predictions of COPD from chest X-ray (CXR) images using deep learning models (Xception and ResNet50). SHAP was specifically employed to provide feature importance visualization and explain model decisions at both global and individual sample levels. The authors used SHAP values to demonstrate how different regions and features in CXRs contributed to COPD diagnosis predictions. The visualization helped identify which areas of the X-ray images were most influential in the model's decision-making process, bridging the gap between complex deep-learning predictions and clinical interpretability.

There are future opportunities to enhance the XAI framework lie in expanding its adaptability to varied CXR perspectives and other imaging techniques like computed tomographic (CT) scans and lung ultrasonography, potentially improving its diagnostic proficiency and contributing vitally to proactive COPD identification and improved patient care. Wang et al.⁵⁷ integrated SHAP with machine learning models (particularly CatBoost) to predict COPD risk in smokers using questionnaires and physical examination data. SHAP was used in multiple ways: (1) To provide global feature importance rankings showing that age, CAT scores, and income were top predictors, (2) To visualize feature value distributions and their impact through SHAP value plots, and (3) To provide personalized prediction interpretations for individual cases. The authors combined SHAP with partial dependence plots (PDPs) to offer both global and local interpretability of the model's predictions. However, the study's predictors are only based on questionnaire information. No lung function monitoring data was recorded. More than that, Deep learning methods have not been investigated in this study.

Applications of SHAP in prediction of diabetes

Joseph et al.⁶⁵ used both model-specific (TabNet's attention mechanism) and model-agnostic SHAP for interpretability. SHAP was specifically used for global explanations while

LIME was used for local interpretations. The study found that for the PIDD dataset, insulin was the most influential feature with SHAP value of 0.301; and for the ESDRPD dataset, polyuria was most influential with a SHAP value of 0.206. The authors used SHAP dependency plots to analyze feature interactions and showed how different features like glucose and insulin levels interacted to influence diabetes predictions. A unique aspect was combining SHAP with other XAI approaches for comprehensive interpretability.

Kibria et al.⁶⁴ applied SHAP to understand feature contributions in a CatBoost model for predicting COPD in smokers. SHAP values helped identify that age, CAT score, and gross annual income were the most important predictors. The authors used multiple SHAP visualization techniques including summary plots, feature importance, and dependence plots to provide both global and local interpretability. The study demonstrated how SHAP can help understand complex feature interactions in both a clinical and lifestyle context.

In another study, the authors implemented LIME and extreme gradient boosting (XGBoost) algorithms with the Highway Network and LeNet model for predicting diabetes.⁶² The authors used SHAP specifically in two novel ways: (1) To provide feature importance visualization of the Hi-Le and HiTCLe ensemble models through waterfall plots showing both local and global interpretations, and (2) to understand the contribution of each feature to individual predictions through SHAP force plots. For instance, with the Hi-Le model, they found that prediabetes and insulin levels were key predictors. What made this paper's use of SHAP unique was its application to ensemble models and the combination of both local and global interpretations to provide comprehensive model explanations. The study also used SHAP alongside other visualization techniques like correlation heatmaps to provide multiple perspectives on feature relationships.

Ganguly and Singh⁶⁶ used SHAP alongside LIME for explainable diabetes prediction. The study focused on using SHAP, particularly for global feature importance and

	Risk of bias						Overall
	D1	D2	D3	D4	D5	D6	
Ahmad et al (2022)	-	+	-	?	?	?	+
Arumugham et al (2023)	-	+	+	?	+	?	+
Baweja et al (2023)	-	+	+	?	+	?	+
Benmohammed et al (2022)	+	+	+	+	+	?	+
Brakefield et al (2022)	-	+	+	?	+	?	+
Choi and Hongchul Lee (2023)	+	+	+	-	+	?	+
Davagdorj et al (2021)	+	+	+	+	+	?	+
Deo and Suranjan Panigrahi (2022)	-	?	+	?	+	?	+
El-Magd et al (2024)	-	+	+	?	-	?	+
Ganguly and Dharmpal Singh (2023)	-	+	+	?	+	?	+
Ghosh and Ahsan H. Khandoker (2024)	-	+	+	+	+	?	+
Hamed et al (2023)	-	+	+	+	+	?	+
Ikechukwu et al (2023)a	+	+	+	?	+	?	+

Figure 11. (A) Traffic-light plot of the risk of bias (RoB) from observational studies.⁹³ (B) Traffic-light plot of the RoB from observational studies.⁹³ (C) Traffic-light plot of the RoB from observational studies.⁹³ (continued)

visualization of feature effects through mean SHAP value plots. The authors used SHAP specifically to understand the impact of plasma glucose levels, which they found to be the most influential predictor.

Applications of SHAP in the prediction of lung diseases

In our systematic review, SHAP is not used to predict, diagnose, treat, or manage any lung diseases. We believe that it might be due to the following reasons:

1. Data type limitations

Most lung disease diagnosis relies heavily on imaging data (CXRs, CT scans) rather than tabular data. SHAP was originally designed and works best with structured tabular data. For image data, other XAI techniques like Grad-CAM and attention mechanisms are more naturally suited for visual explanations.

2. Workflow alignment

Radiologists and pulmonologists are accustomed to visual interpretations. Visual XAI methods like heatmaps directly align with how clinicians already analyze medical

images. SHAP’s feature importance plots may be less intuitive for image-based diagnosis.

3. Technical complexity

Applying SHAP to image data requires sophisticated adaptations. Computing SHAP values for high-dimensional image data can be computationally expensive. Simpler alternatives like Grad-CAM offer sufficient explanations with less complexity.

Other applications of SHAP in the prediction of chronic diseases

Other than the diseases that are mentioned earlier, SHAP is being used to predict, manage, and treat other chronic diseases. Here, we are going through these studies:

Applications of SHAP in the prediction of Alzheimer’s

Sekaran et al.³⁷ employed SHAP as a key technique within an XAI framework to interpret the predictions of machine learning models designed to identify gene biomarkers for Alzheimer’s disease (AD). The primary objective of using SHAP is to provide transparency and biological relevance to the models by explaining which genes significantly

Islam et al (2022)	+	-	+	?	+	?	+
Jhumka et al (2023)a	+	+	+	+	+	?	+
Joseph et al (2022)	+	+	+	?	+	?	+
Junaaid et al (2023)	+	+	+	-	+	?	+
Kamal, Md Sarwar, et al (2023)	-	?	-	?	-	?	-
Kibria et al (2022)	+	+	+	?	+	?	+
Koyyada and Thipendra P. Singh (2023)	-	+	+	?	+	?	+
Li et al (2022)	+	+	+	?	+	?	+
Lo et al (2024)	-	+	+	?	+	?	+
Maouche et al (2023)	-	+	+	?	+	?	+
Morabito, Fortunato, et al (2023)	+	-	+	+	+	?	+
Moreno-Sanchez (2021)	-	+	+	?	+	?	+
Moreno-Sánchez (2023)	+	-	+	?	+	?	+
Moreno-Sánchez et al (2024)	+	-	+	+	+	?	+
Narteni et al (2024)	-	?	+	?	+	?	+

Figure 11. Continued.

contribute to the differentiation between AD and non-AD samples. SHAP is applied to machine learning classifiers, such as logistic regression (LR) and others, that were trained on subsets of differentially expressed genes extracted from datasets. By calculating Shapley values, SHAP quantifies the marginal contribution of each gene to the model's predictions. This approach allows researchers to understand the role of individual genes in influencing the prediction outcome. SHAP visualizations are employed to illustrate the contributions of individual genes to the model's decisions. These visual aids make it easier to interpret how changes in gene expression influence the likelihood of AD. For example, SHAP scores demonstrate the significance of ORAI2 in predicting AD outcomes, reinforcing its potential as a therapeutic target. Such detailed explanations enhance confidence in the model's outputs and provide valuable insights into the biological mechanisms underlying the disease. To further improve this research, external validation is required.

Application of SHAP in the prediction of hematological malignancies

Rodríguez-Belenguer et al.⁷⁰ used SHAP as a critical component of an XAI approach to interpreting a machine learning model

predicting the serological response of patients with hematological malignancies (HMs) to COVID-19 vaccination. The study aims to address the challenge of combining predictive accuracy with clinical interpretability, helping healthcare professionals understand the key factors influencing antibody generation in this vulnerable population. SHAP values are calculated for the best-performing machine learning model, a support vector machine (SVM). These values measure the contribution of each variable to the model's predictions, indicating whether a factor has a positive or negative impact on the likelihood of generating an antibody response. By combining SHAP, principal component analysis, and clustering, the study provides a comprehensive and interpretable methodology for understanding patient profiles and identifying subgroups at higher risk of poor serological response. The results have practical applications, such as informing personalized preventive strategies for HM patients, including additional vaccine doses, prophylactic monoclonal antibodies, or early antiviral therapies.

Application of SHAP in the prediction of liver cirrhosis

Arya et al.⁷⁴ integrated SHAP with an XGBoost classifier to enhance the understanding of the model's predictions, offering transparency and insights critical for clinical

Paul and Md Al Mamun (2023)	-	+	-	+	+	?	+
Pezoulas et al (2022)	-	?	?	-	+	?	+
Prajapati et al (2023)	-	+	+	?	-	?	+
Rodriguez-Belenguer et al (2024)	+	+	+	-	+	?	+
Sekaran et al (2023)	-	+	+	?	+	?	+
Sharma et al (2024)	-	+	+	?	+	?	+
Uegami et al (2022)	-	+	+	?	+	?	+
Ukwuoma et al (2023)	+	+	+	?	+	?	+
Vaccari et al (2021)	-	?	+	+	+	?	+
Vásquez-Morales et al (2019)	-	+	+	?	+	?	+
Vijayvargiya et al (2023)	-	?	+	+	+	?	+
Wu et al (2023)	+	-	+	?	+	?	+
Yagin et al (2023)	-	+	+	?	+	?	+
Yagin et al (2024)	-	+	+	?	+	?	+
Arya et al (2023)	-	+	+	?	+	?	+
Wang et al (2023)	+	-	+	+	+	?	+

Figure 11. Continued.

		Risk of bias						
		D1	D2	D3	D4	D5	D6	Overall
Study	Shang, Yong, et al (2021)	-	?	-	+	+	?	+
	Ikechukwu et al (2023)b	-	+	+	?	+	?	+
	Marvin and Alam (2022)	-	+	+	?	+	?	+
	Huang et al (2024)	-	+	+	?	+	?	+
		D1: Selection bias D2: Performance bias D3: Detection bias D4: Attrition bias D5: Reporting bias D6: Other bias						Judgement - Unclear + Low ? No information

Figure 12. Traffic-light plot of the risk of bias (RoB) from quasi-experimental studies.⁹³

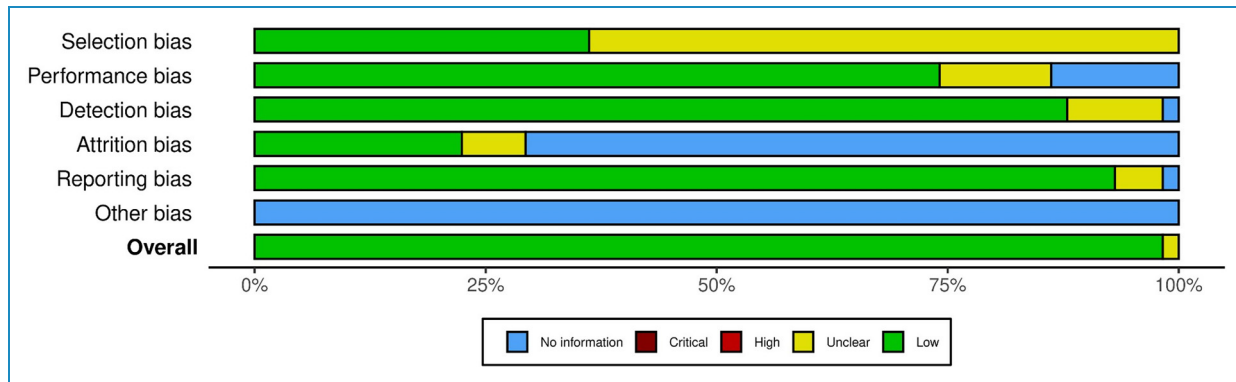


Figure 13. Summary plot of the risk of bias (RoB) from all of the included studies.⁹³

applications. SHAP is used to calculate the contribution of individual features (e.g., platelets, bilirubin levels, cholesterol) to the prediction of liver cirrhosis stages. By assigning Shapley values, the study quantifies how much each feature influences the model's output, highlighting the most critical biomarkers. Moreover, SHAP also evaluates interactions between features by computing the SHAP interaction values. These values capture how pairs of features jointly contribute to the predictions, providing deeper insights into the interdependencies among biomarkers.

It is worth noting that, visual representations of SHAP values, including feature importance plots and feature contribution graphs, are generated. These visualizations clarify which features significantly increase or decrease the likelihood of cirrhosis in the predictions. For instance, cholesterol was identified as having a substantial influence on certain predictions, even though platelets were highlighted as the most important feature overall.

Another aspect of using SHAP that is utilized in this paper is a case-by-case analysis of predictions, explaining how specific feature values lead to a particular output. For example, it shows the magnitude of contributions from features like age or platelets for individual patients.

Application of SHAP in the prediction and management of lymphomagenesis

Pezoulas et al.⁷⁹ utilized SHAP to enhance the interpretability of a federated AI platform developed for disease management, particularly in modeling lymphomagenesis (the development of lymphoma) among patients with rare autoimmune diseases.

Applications of SHAP in prediction of myalgic encephalomyelitis/chronic fatigue syndrome

SHAP is employed by the authors as part of an XAI framework to provide interpretability and transparency to the predictive model designed for classifying myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) patients.⁸³ SHAP is

particularly used to analyze and explain the contributions of individual metabolites to the model's predictions, thereby making the complex machine-learning results understandable for clinicians and researchers. SHAP is applied to the random forest classifier (RFC) trained on biomarker metabolites identified through feature selection. The SHAP analysis reveals the importance of each metabolite in the context of ME/CFS diagnosis by quantifying how much each feature contributes to the predicted outcome. SHAP visualizations highlight the relationship between the values of metabolites and their impact on the prediction. Positive SHAP values signify features that increase the likelihood of ME/CFS, while negative values suggest a protective or neutral role. These graphical outputs not only aid in understanding the model's reasoning but also provide insights into potential biological mechanisms underlying the disease.

Besides the need for external validation, Transfer learning, and deep learning can also be explored in future work.

Yagin et al.⁸² applied SHAP to explain the decision-making process of the XGBoost classifier, which achieved the highest predictive accuracy among the tested models. Using SHAP, the authors identified which metabolites were most influential in predicting ME/CFS and how these metabolites impacted the model's output. To visualize these insights, SHAP summary plots and beeswarm plots were generated. These plots illustrated the contributions of individual features (metabolites) to the model's predictions, which resulted in providing an intuitive understanding of the relationship between metabolite levels and disease risk.

The dataset presents significant limitations due to its imbalanced structure, containing an extensive number of features but insufficient sample size. Furthermore, the absence of critical user demographic information compromises the dataset's completeness and analytical potential.

Application of SHAP in the prediction and management of obesity

Brakefield et al.⁸⁵ utilized SHAP to interpret the predictions of a machine learning model developed within the Urban

Population Health Observatory (UPHO) system. The system is designed for obesity prevalence prediction and decision support in urban population health. SHAP is applied to analyze the contributions of various SDoH features, such as lack of physical activity, poverty, and unemployment, to obesity prevalence predictions. For example, it identifies “lack of physical activity” and “poverty” as the most significant positive contributors to obesity rates in specific neighborhoods. By quantifying the impact of individual features, SHAP assigns importance scores to predictors, helping researchers and public health officials prioritize intervention strategies. SHAP value plots visually demonstrate the direction and magnitude of each feature’s contribution to the model’s predictions. This helps stakeholders, including clinicians and researchers, understand why the model predicts higher or lower obesity prevalence for a given population or region. SHAP is also utilized to provide support for decision-making. By integrating SHAP into the UPHO system, the platform provides actionable explanations through a user-friendly dashboard. These explanations allow physicians and researchers to trace causal pathways linking SDoH to health outcomes like obesity, aiding in the design of more targeted interventions and policies. However, in general, using population data has the limitation of generalizations for individual assessments that are not always necessarily truthful in reality.

Application of SHAP in the prediction of Parkinson’s

Junaid et al.⁸⁷ utilized the SHAP framework to enhance the explainability of machine learning models used for early detection and progression prediction of Parkinson’s disease. SHAP assigns an importance value to each feature, indicating its contribution to a specific decision. This aids in identifying the most critical features influencing the models’ outcomes. A key limitation of this study is the exclusion of ensemble methods and deep learning models from the analysis.

Application of SHAP in the prediction of vascular wound images

Lo et al.⁹⁰ utilized SHAP to evaluate the explainability of the developed AI model for analyzing vascular wound images. SHAP values are assigned to individual pixels of the wound images to measure their contributions to the model’s predictions. Positive SHAP values indicate a pixel positively contributes to the prediction of a specific wound class, while negative values indicate a lack of contribution or a negative impact. The explainability score is derived by combining SHAP values with the proximity of each pixel to the wound area. Pixels closer to the wound with positive SHAP values contribute more significantly to the score. This method ensures the model focuses on relevant areas (wound regions) rather than unrelated parts of the

image. SHAP values are visualized to demonstrate which regions of the wound image the model focuses on when making predictions. Regions with positive SHAP values are highlighted as contributing to the decision-making process. A primary limitation of this study is the model’s performance constraints due to limited data availability and insufficient sample size.

Applications of LIME in the prediction of chronic kidney disease (CKD)

Vijayvargiya et al.⁴⁴ implemented LIME to provide interpretability for a RFC that predicts CKD. LIME is used to explain the model’s predictions by generating local surrogate models that approximate the RF classifier’s behavior for individual predictions. The researchers applied LIME to analyze feature importance and provide explanations for why specific patients are classified as having CKD or not. They presented visualizations showing how different features like creatinine, HgbA1C, and blood pressure contribute to individual predictions, making the model’s decision-making process transparent to healthcare providers.

Arumugham et al.⁵³ employed LIME as an explainability technique for a DNN model that predicts early-stage CKD. The researchers use LIME to generate local interpretable explanations by perturbing input features and observing how they affect the model’s predictions. LIME creates simplified linear models around individual predictions to explain which features most strongly influenced the DNN’s classification decisions. The paper demonstrates LIME’s application through case studies where it explains predictions for specific patients by showing feature importance scores and their contributions (positive or negative) to the prediction. The researchers also use LIME’s explanations to validate that their model is making clinically relevant decisions based on meaningful features rather than artifacts in the data.

Ghosh and Ahsan⁵¹ utilized LIME along with SHAP to explain predictions from an XGBoost model for CKD prediction. LIME is specifically employed to provide instance-level explanations by creating locally faithful linear models around individual predictions. The researchers demonstrate LIME’s application through patient case studies, where they analyze how features like creatinine, glycosylated hemoglobin type A1C (HgbA1C), and age influence specific predictions. The paper shows how LIME generates human-interpretable explanations by identifying which features contribute positively or negatively to each prediction and quantifying their relative importance, helping clinicians understand the reasoning behind individual diagnostic predictions. This study has two notable limitations: the potential lack of generalizability in the dataset and the unexplored application of deep learning methodologies. Islam et al.⁴⁵ employed LIME to provide interpretability for multiple machine learning models, with a

focus on understanding XGBoost's predictions for CKD. LIME is implemented as a model-agnostic explainer that creates locally faithful linear models by perturbing the input data around specific instances. The authors use LIME in several key ways. In terms of Individual Prediction Explanation, LIME enhances model interpretability by generating detailed explanations for individual patient cases, demonstrating how each feature influences the specific prediction. The system visualizes the model's confidence through probability scores, while feature contributions are clearly displayed via bar charts that indicate whether each factor has a positive or negative influence on the predictions.

Regarding Local Feature Importance, LIME provides valuable insights by identifying the most significant features for specific predictions. For each individual case, the system generates a comprehensive ranking of features based on their contribution to that particular prediction. These explanations are presented in an accessible format that clearly demonstrates whether specific features influenced the prediction toward CKD or non-CKD classification. The Clinical Validation aspect demonstrates how LIME's explanations serve to validate the alignment between model decisions and established clinical knowledge. The researchers effectively show how these explanations enable clinicians to better understand and develop trust in the model's predictions. Importantly, the interpretations are presented in a format that resonates with medical professionals, making the system particularly valuable in clinical settings. Figure 4 have utilized LIME to detect, manage, predict, and diagnose chronic diseases. Some of the key trends in this paper are discussed as follows.

In terms of Feature Importance Visualization, Vijayvargiya et al.⁴⁴, Arumugham et al.⁵³, Ghosh and Ahsan⁵¹ and Islam et al.⁴⁵ employed LIME to effectively visualize how different features contribute to individual predictions. The results are consistently presented through bar charts that illustrate positive and negative feature influences. To enhance clarity, feature importance is typically represented through color coding, such as blue and orange, which helps distinguish between factors that support or oppose a CKD diagnosis.

The Individual Case Analysis approach is evident across all papers, with each demonstrating LIME's capabilities through specific patient case studies. These analyses prioritize instance-level explanations over broader model interpretability, placing particular emphasis on local feature importance rather than global feature significance. This focused approach allows for more detailed and relevant insights into individual cases.

Regarding Integration with High-Performance Models, LIME is consistently utilized to provide explanations for sophisticated modeling approaches, including RF, XGBoost, and DNNs. The studies demonstrate LIME's effectiveness in explaining models that achieve high accuracy rates exceeding 90%. This integration serves to justify the implementation of complex models in clinical environments, where interpretability is crucial.

The Binary Classification Focus is maintained throughout all papers, with LIME being applied to explain binary predictions distinguishing between CKD and non-CKD cases. The studies present prediction probabilities alongside detailed feature explanations, with interpretations specifically highlighting the factors that influence classification in either direction. This approach provides clear insights into the decision-making process for binary classification scenarios.

Applications of LIME in the prediction of chronic obstructive lung disease

In our systematic review, LIME is not used for COPD prediction.

Here are several probable reasons why LIME may not be commonly used for COPD prediction, diagnosis, or management based on current medical and technical considerations:

Regarding Disease Characteristics and Diagnosis, COPD presents distinct features that influence its diagnostic approach. The condition primarily relies on spirometry tests and direct lung function measurements for diagnosis, with well-established diagnostic criteria based on objective physiological measurements. Visual diagnostic tools, particularly CT scans and X-rays, play a fundamental role in COPD assessment. Given these clear diagnostic parameters, there may be reduced necessity for complex machine learning models requiring detailed explainability.

The Data Types and Complexity associated with COPD create unique challenges for interpretability tools. COPD data typically encompasses temporal measurements of lung function, complex imaging data from CXRs and CT scans, and detailed waveform data from spirometry tests. LIME's capabilities may be less effective when applied to these complex, multimodal data types compared to its performance with the simpler tabular data structures commonly used in CKD prediction.

Clinical Practice Patterns in COPD management follow well-established protocols. Healthcare providers primarily depend on direct physical examination and spirometry results, with treatment decisions largely guided by protocol-based disease staging. This standardized approach may reduce the immediate need for machine learning decision support compared to conditions like CKD, where decision-making patterns may be more variable.

The Technical Limitations of LIME present significant challenges in COPD analysis. While LIME demonstrates optimal performance with tabular data and straightforward feature sets, it may be less suitable for explaining predictions based on the complex data types characteristic of COPD. These challenges become particularly apparent when dealing with time series data from continuous monitoring, complex imaging features, and scenarios requiring the integration of multiple data modalities.

Applications of LIME in the prediction of diabetes

Nagaraj et al.⁶³ utilized LIME as part of a hybrid approach combining prediction and recommendation for diabetes management. The researchers apply LIME to explain the predictions of their RFC model that processes the PIMA Indians Diabetes Dataset. LIME is specifically used to generate local explanations for individual predictions by showing which features (like glucose levels, body mass index (BMI), etc.) contribute positively or negatively to the diabetes diagnosis prediction. The paper shows LIME's application through visualization tools that help healthcare providers understand why specific predictions were made for individual patients.

Ganguly and Singh⁶⁶ employed LIME alongside SHAP in a complementary approach for diabetes prediction interpretation. LIME is specifically used when both feature importance and anchors (stable explanations) are required simultaneously. The researchers utilize LIME's tabular explainer functionality to provide local interpretations of their ensemble model's predictions. The paper demonstrates LIME's application through class-specific explanations (for both diabetes positive and negative cases) and shows how LIME calculates local model intercepts and prediction probabilities to explain individual cases. LIME is integrated into its web-based interface to provide real-time explanations of predictions to healthcare providers.

Joseph et al.⁶⁵ used LIME as a model-agnostic method to provide local interpretability for their BO-TabNet model's diabetes predictions. LIME was used to explain individual instance predictions by showing how different features contributed to each specific classification decision. For example, when analyzing instance $i=58$ from their first dataset (PIDD), LIME revealed there was a 73% probability of a non-diabetic outcome, with the top five influential features being insulin, glucose, skin thickness, age, and pregnancy. For their second dataset (ESDRPD), LIME showed how categorical features like polyuria and polydipsia strongly influenced the predictions, with values of polyuria >0 and polydipsia >0 favoring diabetic outcomes. The authors highlighted that LIME's advantage was its ability to simply reveal each feature's impact on the outcome shown as a probability, helping make the model's decision-making process more transparent and trustworthy.

Kibria et al.⁶⁴ employed LIME alongside SHAP to explain and visualize the model's predictions, particularly for understanding local explanations of individual cases. The paper used LIME to explain model decisions for specific test samples, especially those with contradictory symptoms that were challenging to predict. LIME generated tabular explanations showing how each feature (like glucose levels, BMI, age, etc.) contributed positively or negatively to diabetes prediction for individual cases. The visualizations presented feature contributions using color coding (blue for negative impact on diabetes prediction,

orange for positive impact) and included probability scores for both classes (diabetic vs non-diabetic). This helped physicians understand why the model made specific predictions on a case-by-case basis. The authors noted that LIME helped make the model's decision-making process more transparent and interpretable for medical professionals.

Applications of LIME in the prediction of lung disease

LIME is utilized as part of a two-phase approach to identify lung nodules in CXR images in research by Koyyada and Singh.⁷⁶ Specifically, after training an initial CNN model, LIME is used as an XAI method to extract regions of interest and local discriminative features from the CXRs. The paper explains that LIME works by dividing the image into sub-regions and feeding each sub-region to the trained model to get predictions. The sub-regions that lead to correct predictions are then highlighted as regions of interest. The number of these regions is controlled by a feature weight parameter. LIME helps balance between interpretability and local explanations, allowing the model to identify specific areas that are most indicative of lung nodules, similar to how a radiologist would examine an image. This local feature extraction through LIME forms part of what the authors call a "two-stage process"—where the first stage uses a CNN for initial classification, and the second stage uses LIME to identify and focus on the most relevant local areas of the X-ray images. This approach helps make the AI system's decision-making process more transparent and interpretable, while potentially improving diagnostic accuracy by focusing on the most relevant regions of the CXRs.

Other applications of LIME in the prediction of chronic diseases

Application of LIME in the prediction of breast cancer. Maouche et al.⁴⁰ employed LIME to provide patient-level explanations for breast cancer metastasis prediction using a cost-sensitive CatBoost classifier. LIME is employed to provide patient-level explanations for breast cancer metastasis prediction using a cost-sensitive CatBoost classifier. LIME quantifies how individual clinicopathological and treatment-related features influence the predicted outcome. It assigns weights to each feature, showing their relative importance in driving the prediction for a specific patient. The output of LIME is visualized as bar charts where features contributing positively or negatively to metastasis risk are highlighted to let the clinicians understand the rationale behind the model's decisions and assess the impact of various factors, such as chemotherapy, radiotherapy, and tumor characteristics, on the metastasis risk. The problem with the explanations is that they are not clear enough and include ML concepts which might be problematic for clinicians to understand.

Application of LIME in the prediction of Parkinson's disease.

Junaïd et al.⁸⁷ used LIME as one of the explainability tools to interpret the predictions made by machine learning models for Parkinson's disease progression detection. The study integrates LIME to enhance the transparency of the models, allowing both global and local explanations for the predictions, which is critical in medical applications. LIME provides local explanations by approximating the predictions of complex models with simpler interpretable surrogate models. These explanations are used to understand the behavior of classifiers, such as RF, Extra Trees classifier (ETC), and light gradient boosting machines (LGBM), in specific instances of the dataset. The paper uses LIME alongside other explainability methods, such as SHAP and Shapash, to ensure the consistency and reliability of explanations. LIME's outputs are compared to these methods to confirm its validity in identifying critical features influencing model predictions. LIME's outputs are visualized to show the contribution of each feature to a specific prediction. A key limitation of this study is the exclusion of ensemble methods and deep learning models from the analysis.

Application of LIME in the prediction of sleep apnea.

Troncoso-Garcia et al.⁸⁶ utilized LIME to provide interpretability for machine learning models applied to the detection of sleep apnea using data from polysomnography (PSG). LIME is applied to explain the predictions of the RFC, which was identified as the best-performing model in terms of accuracy and other quality metrics. It provides local explanations for individual predictions by analyzing the contribution of specific features (attributes of the PSG signals) to the model's classification decision for each instance. LIME perturbs the input data around a specific instance and observes the changes in the model's predictions. This allows the identification of the most influential features and their respective values for classifying a particular instance as an apnea event (class 1) or a non-apnea event (class 0). The study shows that critical features for prediction often relate to abrupt changes in nasal airflow, which are indicative of sleep apnea. To make the results more intuitive for clinicians, the study also integrates LIME explanations with the graphical representation of airflow signals. Red points on these graphs mark the critical features identified by LIME. This way, it makes it easier for health professionals to understand how specific points in the signal influence predictions. A key limitation of this research is the absence of deep learning methodologies in the analysis framework.

Applications of grad-cam in the prediction of CKD. Mehedi et al.,⁴⁶ utilized Grad-CAM as an XAI technique to provide visual explanations for the VGG16 model's kidney tumor classification predictions. As shown in Figure 9 of the

paper, Grad-Cam helps visualize which specific features the deep learning model focuses on when making its classification decision. The authors applied Grad-Cam to four sample images (two normal and two tumor images) to demonstrate how the model selects relevant features for accurate classification. For instance, in the normal kidney samples, Grad-Cam highlights the areas that the model considers most important in determining the image as normal, while for tumor samples, it shows the regions that contribute to identifying the presence of a tumor. This technique adds interpretability to the black-box nature of DNNs by revealing the key visual cues that drive the model's decision-making process, thereby increasing transparency and helping researchers understand how the machine-learning model arrives at its predictions.

Applications of grad-cam in prediction of COPD. Ikechukwu et al.⁵⁵ utilized Grad-CAM as an XAI technique to provide visual explanations for the deep learning model's predictions. Specifically, the authors use Grad-CAM to highlight the most important regions in CXR images that contribute to the model's classification of COPD-related conditions. In the study, Grad-CAM helps to:

1. Visualize which specific areas of the CXR images are most significant for the model's decision-making process.
2. Provide transparency to the AI model's predictions by showing the key features that influenced the classification.
3. Enhance the interpretability of the deep learning model for healthcare practitioners by demonstrating how the model arrives at its diagnosis.

Future opportunities to enhance the XAI frameworks include expanding its adaptability to diverse chest X-ray (CXR) views and incorporating other imaging modalities such as computed tomography (CT) scans and lung ultrasonography. These advancements have the potential to significantly improve diagnostic accuracy and play a critical role in the early detection of COPD, ultimately leading to better patient outcomes and more proactive care.

In another study done by Ikechukwu et al.,⁵⁶ Grad-CAM is utilized as an explainability method to generate visual explanations for the model's predictions. Specifically, the authors employ Grad-CAM to highlight the regions in CXR images that significantly contribute to the model's diagnostic decisions. By generating heat maps that indicate the most important areas influencing the prediction, Grad-CAM helps clinicians understand and trust the model's outputs. As described in the paper, this visual explanation technique is crucial in healthcare applications, where interpretability and confidence in AI-driven diagnostic models are paramount.

However, it is crucial to consider other metrics such as specificity and AUC ROC to evaluate these models'

performance fully. Although CT scans are considered the “gold standard” for capturing lung regions, it is not clear how the proposed model generalizes to real-time, heterogeneous data.

Applications of Grad-CAM in prediction of diabetes. In our systematic review, Grad-Cam is **not** used to predict diabetes. There could be many reasons Why the Grad-Cam is not a suitable choice for predicting Diabetes. Some of the main reasons, however, could include the following:

1. **Data Complexity:** Diabetes diagnosis might rely more on complex multi-parameter clinical data rather than visual imagery, which is where Grad-CAM typically excels.
2. **Limited Imaging Data:** Unlike diseases like COPD or COVID-19 where CXRs provide clear visual indicators, diabetes diagnosis traditionally depends more on blood tests, glucose levels, and clinical parameters.
3. **Model Interpretability Challenges:** Some diabetes prediction models might use complex ensemble or multi-feature machine learning techniques that are not easily interpretable with visual techniques like Grad-CAM.

Applications of Grad-CAM in the prediction of lung disease.

Grad-CAM is used as an XAI technique to provide visual explanations for the model’s predictions in a study done by Sharma et al.⁷⁵ Specifically, the authors incorporate Grad-CAM to highlight the infected areas in the CXR images, allowing both doctors and users to understand precisely why a particular classification was made. The Grad-CAM model generates a visual heatmap that overlays the original X-ray image, emphasizing the regions that most significantly contributed to the pneumonia classification. This approach addresses one of the key challenges in medical image classification by providing transparency in the decision-making process. As the authors note in their conclusion, the Grad-CAM visualization helps doctors and even non-expert users identify the specific class of pneumonia and understand the reasoning behind the model’s output by visually pointing out the infection spots in the CXR. Future research requires rigorous clinical validation through real-time implementation in healthcare settings.

In another study by Koyyada and Singh,⁷⁶ Grad-CAM is utilized as part of the XAI approach to enhance the interpretability of lung disease predictions, particularly for identifying lung nodules from CXR images. Grad-CAM generates heatmaps highlighting regions of interest within the X-ray images that strongly influence the CNN model’s predictions. This enables the localization of critical features associated with potential lung nodules, aligning with radiologists’ methods of analyzing such images. The process

involves first training a CNN to predict lung disease, followed by applying Grad-CAM to produce visual explanations that identify discriminative regions within the lungs. These heatmaps guide the refinement of the detection process, making the predictions more transparent and aiding medical practitioners in correlating the AI’s insights with clinical findings. This integration of Grad-CAM not only improves model explainability but also ensures that the identified features have medical relevance, addressing the black-box nature of CNNs.

Other applications of Grad-CAM in the prediction of chronic diseases

Application of Grad-CAM in prediction of knee osteoarthritis.

Ahmed et al.⁷² employed Grad-CAM as part of an XAI framework to interpret the predictions of deep learning models used for diagnosing knee osteoarthritis (OA) from X-ray images. Grad-CAM is utilized to visualize the decision-making process of CNNs by highlighting the specific regions of the knee joint that influence the model’s predictions. Grad-CAM generates heatmaps that show the regions in the X-ray images where the model focuses during classification. This visualization helps determine whether the model considers medically relevant areas, such as the knee joint, when predicting the severity of OA based on the Kellgren–Lawrence (KL) grading system.

The study applies Grad-CAM to both multi-class and binary classification tasks. For the multi-class classification, the Grad-CAM heatmaps revealed that while the model focused on the knee joint region, it often failed to differentiate effectively between certain KL grades (e.g., grades 0 and 1). For binary classifications (e.g., normal vs. severe cases), Grad-CAM showed improved attention to relevant areas, correlating with higher classification accuracy. Grad-CAM was used to analyze the model’s ability to make decisions similar to medical professionals. For example, in severe OA cases (KL grade 4), the heatmaps indicated that the model concentrated on the narrowed joint spaces, which aligns with clinical diagnostic criteria.

Application of Grad-CAM in the prediction of ulcers.

Lo et al.⁹⁰ used Grad-CAM to generate heatmaps that highlight the specific regions in wound images that the model focuses on when making predictions. This is critical in ensuring that the model is paying attention to relevant wound areas, such as wound boundaries and characteristics, rather than irrelevant parts of the image in this study. Grad-CAM contributes to calculating an explainability score by measuring the alignment between the model’s attention and clinically relevant regions of the image. For example, areas close to the wound bed and surrounding tissue are weighted more heavily in this Grad-CAM is used alongside other XAI methods like LIME and SHAP to validate the model’s

interpretability. This multi-method approach ensures a comprehensive understanding of the model's behavior calculation. However, insufficient data remain a significant limitation in model performance.

Application of Grad-CAM in the prediction of leprosy. Baweja et al.⁷³ employed Grad-CAM to generate heatmaps that identify the regions in an image most important for the model's classification. These heatmaps allow visualization of the areas the model considers when predicting whether an image contains leprosy lesions or not. For example, Grad-CAM demonstrates that LeprosyNet accurately focuses on lesion-affected areas in the image, highlighting critical features such as the periphery and texture of the lesion. More importantly, Grad-CAM is used to validate the interpretability and robustness of LeprosyNet's predictions. By visually confirming that the model consistently focuses on lesion-relevant regions, the study underscores the reliability of the predictions.

Other XAI algorithms in the prediction of chronic diseases

Case-based reasoning in prediction of COPD. El-Magd et al.⁵⁹ employed case-based reasoning (CBR) model as part of a twin system with a deep learning model (DNN) to provide interpretable explanations for the classification of COPD using exhaled breath data. The CBR model utilizes feature weights derived from the DNN using a Hadamard product-based method called contributions-oriented local explanations (C-HP). This method highlights the features most significant to the DNN's predictions. The CBR model explains the DNN's predictions by retrieving similar cases from the training dataset based on the feature weights. This allows for intuitive reasoning by presenting precedents (similar cases) that support the classification decision. The agreement between the DNN and the CBR model is used as a metric for fidelity, ensuring that the explanations closely reflect the DNN's decision-making process.

Pixel density analysis and probabilistic graphical model in the prediction of Alzheimer's. Kamal et al.³⁶ used pixel density analysis (PDA) and probabilistic graphical models (PGMs) to provide explainability in detecting cerebral microbleeds (CMBs) and analyzing AD severity using multimodal data. PDA identifies CMBs by pinpointing brain regions with abnormal pixel density. For example, in MRI images with AD, PDA highlights microbleed regions using red color boxes and maps areas of low pixel density indicative of brain abnormalities. PGM is applied to gene expression data to identify relevant genes associated with AD. PGM analyzes correlations among genes and identifies biological dependencies to explain which genes are most relevant for AD. The study finds specific genes

such as CTAGE6, F8A2, and SAMD7 as strongly associated with AD based on probabilistic dependencies. PGM provides an interpretable explanation of which genes play critical roles in the progression and severity of AD, enabling better understanding for medical professionals.

Most-Weighted path, most-weighted combination, and Maximum frequency difference explanation in prediction of diabetes. The authors⁶¹ used three distinct XAI techniques to enhance the interpretability of diabetes prediction models. The Most-Weighted Path Explanation traces back from the output layer to identify the single most influential input feature by analyzing neural network weights. This provides straightforward explanations such as linking elevated glucose levels to diabetes prediction. The most-weighted combination explanation examines pairs of input features, utilizing a nearest-neighbor approach to identify the two most significant features that work together to influence the prediction. This method specifically focuses on understanding how different features interact to contribute to the diagnosis. The maximum frequency difference explanation compares the occurrence rates of input parameters between diabetic and non-diabetic cases. This technique calculates frequency differences to identify which parameters show the most significant statistical variation between the two groups, thereby highlighting the most discriminative features for diabetes detection.

Additive explanatory factors (explainable boosting machine produced) in the management of heart disease/cardiovascular disease/lung disease. Pfeuffer⁶⁹ utilized additive explanatory factors (explainable boosting machine (EBM)-Produced) as part of an EBM to provide intelligible and interpretable explanations in a patient education system for cardiovascular disease risk. The system uses EBMs, a variant of generalized additive models, which are inherently interpretable and perform well on medical datasets. EBMs produce explanations by breaking down the model's predictions into individual contributions from each feature (health indicator). These additive contributions (explanatory factors) are displayed as importance bars, highlighting features that increase or decrease the risk of cardiovascular disease. On the patient side, they can interact with the system to explore how changes in their health indicators affect risk predictions. If a user excludes certain examination results, the system updates the additive explanatory factors and recalculates the model's predictions. Patients can compare their results with others, enabling contrastive learning to better understand their health status. This research requires rigorous clinical validation and continuous adherence to established medical standards to ensure its practical applicability in healthcare settings.

Layer-wise relevance propagation in prediction of COPD. Ikechukwu et al.⁵⁶ used layer-wise relevance propagation

(LRP) as an explainability technique to provide insights into the decision-making process of the proposed COPDNet model, which is based on the ResNet50 architecture for diagnosing COPD from CXR images. The authors used LRP to generate visual explanations by identifying which regions of the CXR images contribute most to the model's predictions. These explanations highlight the areas that the model considers most relevant for classifying COPD, making the predictions transparent. Therefore, LRP improves the interpretability and trustworthiness of the model, particularly for clinicians who need to understand the basis of the predictions. LRP is used alongside Grad-CAM (another explainability technique), providing multiple perspectives on how the model arrives at its conclusions. While Grad-CAM generates coarse localization maps, LRP provides a more detailed pixel-level explanation.

Graph properties in prediction of CKD. Shang et al.⁴⁸ used graph properties as part of an EHR-oriented knowledge graph system designed to efficiently utilize previously neglected information in electronic health records (EHR). The system converts structured EHR data into a graph-based knowledge model using resource description framework triples. Each patient's data are modeled as a semantic graph, where entities (e.g., patients, diagnoses, test results) are represented as nodes and relationships (e.g., connections between diseases and test results) are represented as edges. This graph-based framework is further detailed through the following components, which outline its ontology structure, reasoning capabilities, visualization methods, and the personalized organization of patient data.

Ontology-Based Structure:

- A 2-level ontology structure is used to define graph semantics. This includes:
 - A top-level ontology for general medical concepts and relationships (e.g., patients, visits, diagnoses).
 - A disease-specific local ontology that adds specific knowledge about conditions (e.g., CKD).

Graph Reasoning:

- The system uses semantic reasoning to traverse and analyze relationships within the knowledge graph. Rules are applied to identify important patterns (e.g., CKD risks based on test results).
- Graph-based reasoning helps identify regions of interest-specific time windows where abnormal clinical findings occur.

Visualization and explanation:

- The graph properties are leveraged to visualize clinical trajectories and reasoning pathways. Nodes and

edges representing patient data (e.g., diagnoses, test results, and relationships) are plotted over time to create an interpretable clinical timeline.

- Key regions of interest (e.g., abnormal kidney function markers) are highlighted for clinicians.

Eventually, each patient's medical data are organized into a 3-level patient-visit-treatment graph, which supports the following:

- Personalized reasoning: Identifying clinical risks based on graph relationships.
- Trajectory analysis: Mapping the sequence of clinical visits and associated treatments.

A significant limitation of this study is its inadequate implementation of data privacy measures, which raises concerns about the protection of sensitive medical information.

Partial dependency in prediction of hepatic steatosis | non-alcoholic fatty liver disease

Deo and Panigrahi⁷¹ used PDPs to analyze the contributions of physiological and demographic variables in predicting hepatic steatosis (HS) using SVM models. The study focuses on understanding how specific features influence the classification outcomes of HS/no-HS across different models and genders, providing insights to enhance model interpretability and performance. PDP is utilized to examine the effect of individual predictors, such as age, BMI, alanine aminotransferase (ALT), aspartate aminotransferase, and glucose levels, on the SVM model's prediction scores. It highlights how changes in these variables correlate with the likelihood of HS. For instance, an increase in age beyond 50 or a BMI above 30 shows a positive trend in disease risk, as reflected in higher classification scores. In the case of gender-specific trends, separate PDP analyses are performed for male and female populations to account for gender-dependent pathobiology in HS. This approach reveals distinct trends, such as higher classification scores for women compared to men across various variables like ALT and glucose levels, suggesting gender-specific differences in disease expression.

Moreover, PDPs are generated for three SVM models: Quadratic SVM, Gaussian Scale 1, and Gaussian Scale 2. These plots allow comparisons of how each model captures relationships between predictors and outcomes. For example, the Gaussian Scale 1 model effectively captures increasing and plateauing trends in BMI and liver enzyme parameters.

Finally, the study uses PDP to validate the clinical relevance of key predictors, with the authors finding that trends observed in ALT, AST, and glucose levels align with established literature supporting their use as top predictors of HS.

Explanation module in prediction of metabolic syndrome

Benmohammed et al.⁸¹ used the explanation module as a key component of the methodology used to develop AI-based scores for screening metabolic syndrome (MetS) in adolescents. It addresses the challenge of understanding and interpreting the predictions made by AI models, ensuring that the resulting scores are both accurate and clinically relevant. This module is designed to extract meaningful insights from the black-box AI models (e.g., artificial neural networks) by explaining the learned decision functions. This enables the derivation of new, simplified MetS screening scores that rely on a reduced set of clinical and biological variables. The process overview includes three main stages as follows:

1. Data Generation: Artificial data resembling real data are generated to characterize the decision boundaries between MetS-positive and MetS-negative classes.
2. Model Labeling: The trained black-box model assigns labels to the generated data, revealing the behavior of the decision surface.
3. Decision Hyperplane: The explanation module identifies the decision function (a hyperplane) that separates the two classes. The hyperplane is mathematically expressed as a combination of selected variables and coefficients derived from the AI model.

Regarding the feature selection, The module incorporates a two-step feature selection process. First, it uses machine learning techniques (e.g., Gini importance) to automatically rank variables. Then, domain experts (medical doctors) review the results to finalize the most relevant variables. For this study, age, waist circumference, mean blood pressure, and the triglyceride-glucose index were selected as the most critical predictors.

The explanation process extracts coefficients for each feature and uses them to create linear equations representing the AI-based MetS scores. These scores are optimized for simplicity, accuracy, and ease of use in clinical practice.

DeepSHAP in prediction of non-communicable diseases (NCDs)

Davagdorj et al.⁸⁴ employed DeepSHAP to interpret the predictions made by a DNN for the early detection of NCDs. The study integrates DeepSHAP within an XAI framework to provide both global and local-level explanations of the model's behavior.

In the case of Global Explanations (Population-Based Perspective), DeepSHAP is employed to explain the collective influence of features on the model's predictions across the entire population. The framework identifies and

ranks the most significant features based on their importance scores, enabling researchers to understand which risk factors (e.g., high blood pressure, BMI, cholesterol levels) contribute most to NCD predictions at the population level. Features such as "Ever told you had high blood pressure," "Age in years," and "Body Mass Index (BMI)" were identified as the most influential in predicting NCD risk.

In the case of Local explanations (Human-Centered Perspective), DeepSHAP provides instance-level explanations by evaluating how specific features contribute to the prediction for a single individual. This allows the model to explain why a particular prediction was made for a specific patient. For instance, in a randomly selected individual, features like "Body Mass Index (BMI)" and "Education Level" were highlighted as either positively or negatively associated with the prediction outcome. Moreover, the DeepSHAP approach enables a layer-wise propagation of SHAP values, leveraging the DeepLIFT methodology. This allows for efficient computation of Shapley values in complex deep learning architectures like DNNs, making the explanations computationally feasible and interpretable.

Attention mechanism in prediction of pneumonia

Ukwuoma et al.⁸⁸ utilized an attention mechanism through a Transformer Encoder with multi-head self-attention to enhance the classification of pneumonia using CXR images. The attention mechanism addresses the challenges of feature extraction and interpretation, enabling more accurate and explainable predictions in both binary and multi-class pneumonia classification tasks. The main purpose of utilizing the Attention mechanism is that it allows the model to focus on the most informative regions of CXR images, discarding redundant features. This enhances the model's ability to distinguish between healthy and pneumonia-affected areas, even in cases of ambiguous image quality. More than that, unlike conventional CNNs that only analyze local correlations between spatially neighboring pixels, the multi-head self-attention mechanism identifies relationships between distant pixels. This capability is particularly useful for medical images like X-rays, where patterns indicating disease may span across non-adjacent regions. The attention mechanism contributes to generating explainable visual outputs, such as heatmaps, which highlight the regions of the X-ray image that the model considers critical for its predictions.

LRP in the prediction of pulmonary (normal, COVID-19, Tuberculosis, pneumonia)

Marvin et al.⁸⁹ employed LRP to explain the predictions of a deep learning model developed for pediatric pulmonary health evaluation. The model aims to classify diseases such as COVID-19, Pneumonia, Tuberculosis, and normal

cases from CXR images. LRP is used to make the neural network's decisions interpretable by identifying the specific image regions that contributed to each classification result.

LRP provides instance-specific explanations by generating heatmaps that visualize the importance of individual pixels in the CXR images. These heatmaps highlight regions of the X-ray that the model deemed significant for its prediction, such as areas showing inflammation or abnormalities in the lungs. The heatmaps produced by LRP made it clear where the model is focusing when identifying specific pulmonary diseases. The paper compares LRP with other explainability methods, such as occlusion and SmoothGrad, and concludes that LRP provides the most interpretable explanations.

Fuzzy ID3 in prediction of respiration diseases

Li et al.⁹² utilized the Fuzzy ID3 algorithm to build fuzzy decision trees that approximate the predictions of a student neural network model. These trees provide step-by-step reasoning that is easy for physicians to understand. By leveraging fuzzy logic, the Fuzzy ID3 algorithm deals with the inherent uncertainty and variability in medical data.

In the next stage, the constructed fuzzy decision trees underwent a pruning process to reduce complexity and improve generalization.

The fuzzy decision tree has generated interpretable decision rules that map input features to predictions. These rules can be visualized and directly used by clinicians to understand the basis of the model's predictions.

However, the dataset could also be more comprehensive to test if the system is robust against high variability.

Applications of XAI algorithms in the treatment of chronic diseases

Hamed et al.⁵² employed SHAP to enhance the explainability of the deep learning model used to predict medical resource utilization for CKD patients. In this research, SHAP is applied to identify the most influential variables that drive the model's predictions. Variables such as CKD-related metrics (e.g., eGFR, uACR), medical codes (diagnoses, procedures, medications), and non-medical factors (e.g., patient age, gender, and prior care consumption) are assessed for their contribution to the model output. Specifically, SHAP is used to interpret predictions across different temporal aggregation periods (e.g., one, two, four, six months). This reveals how temporal granularity affects the importance of variables and overall model performance. However, it seems necessary to consider longer periods of patient history than what is done in this study. Huang et al.⁶⁸ used SHAP to provide insights into how specific features, such as a patient's medication history or clinical attributes, influence the recommendation outputs which results in helping patients and healthcare professionals trust the system's suggestions. Using SHAP decision plots to illustrate the incremental

impact of each feature on the final prediction, aids in a detailed understanding of the decision-making process.

LIME is also used to provide local explanations for the drug recommendation system's outputs in this paper. It generates perturbed versions of an individual data instance (e.g., a patient's health and medication history), applies the original model to predict outcomes, and fits a simpler surrogate model (like linear regression) to approximate the model's behavior around that instance. LIME produces intuitive visualizations, such as bar charts, that show the weights of individual features affecting a recommendation. By making the system's logic accessible and understandable, LIME helps build trust among users. Healthcare professionals can better understand why certain drugs are recommended, enabling them to validate or challenge the suggestions based on their clinical judgment.

While SHAP offers global and local explanations for the model, LIME focuses on creating locally interpretable models specific to individual recommendations. Together, they provide a comprehensive view of how recommendations are generated.

Applications of XAI algorithms in the management of chronic diseases

Moreno Sanchez et al.⁶⁷ employed PDPs and mean decrease impurity (MDI) as part of the explainability techniques to analyze the predictions of machine learning models developed for classifying fibromyalgia (FM) severity. PDP is applied to visualize the marginal effect of individual features on the probability of predicting specific FM severity levels. For example, the study uses PDP to show how changes in mental health variables (such as anxiety and depression scores) influence the likelihood of being classified into "very severe" FM. For each feature, the PDP graph depicts the feature's range on the horizontal axis and its quantitative contribution to the prediction probability on the vertical axis. PDP highlights global trends across the dataset by showing the average impact of a feature on the model's predictions. For instance, anxiety levels above a specific threshold (e.g., 30) were associated with an increased probability of very severe FM, confirming its clinical relevance. On the other hand, MDI is used to measure the importance of each feature in ensemble tree-based models by quantifying the decrease in node impurity (e.g., Gini index or entropy) each time the feature is used for splitting. Features with higher MDI scores are deemed more critical for the model. In the context of FM severity prediction, MDI ranks mental health factors (such as anxiety and depression) as more important than pain-related factors (e.g., McGill Pain Index, pain intensity). MDI provides a straightforward numerical ranking of features, aligning with clinical expectations. For example, anxiety and depression consistently ranked as the top contributors to FM severity classifications across different target scales (PDS and FIQ). Vaccari et al.⁵⁸ employed the logic learning model (LLM) to generate intelligible rules in the form of "if-then" statements, which makes it easier for clinicians to

understand the relationships between different features and their impact on patient outcomes. These rules are derived by clustering data samples into logical groups. The LLM approach emphasizes extracting rules that are statistically validated. The rules generated are evaluated for their significance using Fisher's exact test, ensuring they are reliable for decision-making. The LLM process includes a feature selection mechanism, identifying the most relevant features contributing to the outcomes while minimizing redundancy. This simplifies the analysis for clinicians by highlighting critical factors influencing disease progression. Furthermore, by transforming data into a Boolean space and clustering it, LLM keeps computational costs manageable, allowing for the processing of large datasets without compromising performance. Eventually, the model specifically focuses on monitoring COPD-related parameters such as Forced Expiratory Volume (FEV1) and Peak Expiratory Flow (PEF). The generated rules help classify patient conditions and inform potential interventions. To enhance the generalizability of synthetic data, future research should expand the study to include a larger participant cohort. Narteni et al.³⁹ utilized logic learning machine (LLM) as an XAI model to predict cough-related QoL impairments in asthmatic patients. The LLM generates if-then rules to classify asthmatic patients into two groups: those with impaired QoL and those with near-normal QoL. These rules provide explicit and interpretable criteria for clinicians. The LLM model is trained on a dataset split into 70% for training and 30% for testing. The rules generated by the LLM are statistically validated using Pearson's Chi-Square test to ensure their significance. The LLM computes a feature ranking to identify which symptoms contribute most to predicting impaired QoL. The ranking highlights key areas for clinical focus as follows:

- For controlled asthma, pharynx/larynx and Rhino-sinusitis are the most significant.
- For uncontrolled asthma: Asthma-related symptoms and Gastroesophageal reflux dominate.

For the decision rules, the LLM generates specific threshold-based rules to identify critical symptom levels associated with impaired QoL. The LLM achieves an accuracy of at least 70% across different patient groups, demonstrating its reliability for clinical use.

Rationale behind the selection of XAI methods for chronic disease applications

SHAP: Dominant in structured clinical data analysis. SHAP was by far the most frequently used XAI method (as shown in Figure 5), reflecting its dominance with structured clinical data. Several key factors contribute to its dominance as in the following:

LIME's appeal in settings that prioritize interpretability over exact precision becomes clearer when viewed through the

lens of real-world deployment challenges. In many low-resource healthcare environments, factors such as limited computational power, low digital literacy, and fast-paced clinical workflows demand explanation methods that are both efficient and easy to understand, as explained in the following:

Research comparing multiple XAI approaches for traumatic brain injury prediction found that "SHAP is the most stable with the highest fidelity" compared to other interpretation techniques.⁹⁷ This stability is crucial in healthcare applications where consistent explanations build clinical trust. In diabetes prediction specifically, SHAP applied to gradient-boosting decision tree models successfully identified blood glucose, BMI, and age as key predictive features that directly aligned with established clinical knowledge.⁹⁸ This validation of known clinical factors enhances credibility among healthcare practitioners and patients.

(2) Feature importance quantification

SHAP's mathematical foundation provides precise quantification of how important each feature contributes to the final prediction, making it particularly valuable for chronic disease classification where explainability is crucial. For hospital data analysis, researchers have developed enhanced SHAP implementations, including "a new metric of feature importance using SHAP and a technique termed feature packing, which packs multiple similar features into one grouped feature to allow an easier understanding of the model".⁹⁹

(3) Granular feature analysis for complex conditions

For chronic conditions requiring detailed monitoring, SHAP offers exceptional granularity. Studies demonstrate that "SHAP can provide granular insights into the input feature contribution to the prediction outcome," which is particularly valuable in "high-stakes applications like healthcare or rehabilitation".¹⁰⁰ Overall, greater emphasis on model transparency, integration of advanced explainability techniques, and standardization of interpretability approaches in healthcare ML applications are a necessity in future research regarding the usage of SHAP in healthcare.¹⁰¹

LIME: Preferred for interpretability and local explanations

LIME has gained significant traction, particularly in contexts where simplified, accessible explanations are prioritized over absolute mathematical precision.

LIME's appeal in settings that prioritize interpretability over exact precision becomes clearer when viewed through the lens of real-world deployment challenges. In many low-resource healthcare environments, factors

such as limited computational power, low digital literacy, and fast-paced clinical workflows demand explanation methods that are both efficient and easy to understand, as explained in the following:

In low-resource healthcare systems, computational constraints significantly influence XAI method selection. LIME's model-agnostic approach requires $\sim 10\times$ fewer computational resources than SHAP for generating explanations, making it more feasible for deployment on legacy hardware commonly found in developing regions.^{102,103} For example, a study on AI adoption in Ethiopian hospitals found that SHAP's exponential time complexity ($O(2^M)$ for M features) often exceeded available compute capacity, while LIME's linear sampling approach ($O(N)$) enabled practical implementation on low-end devices.^{103,104} This efficiency aligns with the findings of research optimizing AI for the Global South, which emphasizes that "excessive computational demands risk excluding frontline health workers from benefiting from AI advancements."¹⁰² Simplified explanations become critical when deploying systems in regions with unreliable power grids or limited cloud infrastructure.

(2) End-user expertise and literacy gaps

In many developing healthcare systems, clinicians and community health workers lack advanced technical training. LIME's feature importance rankings and binary relevance scores (e.g., "Blood glucose level contributed +0.3 to diabetes prediction") prove more actionable than SHAP's complex Shapley value distributions.^{105,106} A 2025 study in Nigeria demonstrated that nurses interpreted LIME explanations correctly 78% of the time versus 42% for SHAP visualizations, due to SHAP's reliance on abstract concepts like baseline expectations and additive feature interactions.¹⁰³ This aligns with WHO recommendations for "cognitively aligned AI explanations" in low-literacy settings, where interpretability must account for variations in clinical training levels and digital fluency.^{102,104}

(3) Integration with existing workflows

Resource-constrained clinics often prioritize minimal workflow disruption. LIME's ability to generate explanations in under 2 s per prediction (vs. SHAP's 10–30 s) makes it compatible with fast-paced outpatient environments.^{107,108} For chronic disease screening camps in rural India, LIME-enabled TB detection systems reduced explanation generation time by 83% compared to SHAP-based alternatives, enabling clinicians to process 50+ patients daily without delays.^{102,103}

(4) Regulatory and ethical considerations

Emerging AI governance frameworks in developing nations increasingly mandate minimum explainability

standards rather than maximal precision. India's 2024 Digital Health Act, for instance, requires "clinically actionable explanations" but does not enforce mathematical rigor.¹⁰⁴ LIME satisfies these requirements while avoiding the technical debt of maintaining SHAP's exact coalitional game theory calculations.¹⁰⁹

(5) Cultural factors in trust building

Qualitative studies in Kenya and Bangladesh reveal that clinicians distrust "overly precise" AI explanations, perceiving them as black boxes in disguise.^{103, 104} LIME's deliberate approximation (using linear surrogates) paradoxically enhances trust by mirroring human heuristic reasoning. As noted in a Malawi HIV clinic trial: "Nurses trusted LIME's 'good enough' explanations precisely because they resembled their own diagnostic shortcuts."¹⁰²

Moreover, a 2024 case study on diabetes management in Pakistan, in which AI screening tools were deployed across 12 rural clinics, compared SHAP and LIME explanations. Using SHAP, the study found 68% clinician adoption rate, with 41% reporting that "explanations created confusion". Moreover, using LIME, it found 92% adoption rate, with 79% stating that explanations "matched their diagnostic logic". The study concluded that LIME's intentional simplicity better accommodated: (1) intermittent electricity (reduced compute needs); (2) multilingual health workers (easier translation of binary relevance scores); and (3) paper-based record systems (concise printouts).

(6) Accessibility for clinical practitioners

LIME is often favored for its ability to provide intuitive explanations that can be readily understood by clinicians. Research evaluating LIME's effectiveness for tabular data demonstrated its utility "in terms of making tabular models more interpretable and how it can "supplement conventional performance assessment methods."¹¹⁰

(7) Application in colorectal cancer research

For complex diseases such as colorectal cancer, LIME has demonstrated particular value. For instance, Lee et al. applied it to interpret a Transformer architecture model, which provided explanations for both local and global level predictions that foster the understanding of the model's decision-making process and offer healthcare practitioners valuable insights.¹¹¹

(8) Complementary use with other methods

LIME is frequently employed alongside other XAI techniques for a comprehensive understanding of the decision-making process of AI models. For example, a

comparative study that analyzed cancer patient data utilized both SHAP and LIME along with a patient case study, which gave an accurate and trustworthy result. Regarding medical imaging analysis, Grad-CAM, an XAI technique used to understand the parts of an input image that are most important for CNN classification decision-making, has emerged as the preferred method for medical imaging applications across various chronic diseases.

(1) Visual interpretation strengths

While less frequently used overall (12.0% of applications), Grad-CAM shows particular strength in medical imaging contexts. This specialization makes it invaluable for diseases where diagnostic imaging plays a crucial role, including various cancers, cardiovascular conditions, and neurological disorders.

Disease-specific XAI method selection patterns. The choice between XAI methods appears to be influenced by specific disease characteristics, data modalities, and intended applications.

(1) Cardiovascular disease prediction

For cardiovascular disease prediction, SHAP has emerged as particularly valuable due to its ability to handle complex interactions between multiple risk factors. Its mathematical foundation aligns well with the multifactorial nature of cardiovascular disease risk assessment.

(2) Diabetes management

In diabetes applications, SHAP has demonstrated exceptional utility, particularly because “blood glucose, BMI, and age were identified as key predictive features, aligning with clinical knowledge.” Furthermore, “SHAP explanations improve the model’s explainability and credibility, supporting personalized treatment”.⁹⁸

(3) Cancer diagnostics and prognosis

Cancer applications show a more diverse approach to XAI method selection, reflecting the complexity of oncological data: For structured data analysis (e.g., biomarkers, patient history), both SHAP and LIME are commonly employed. For imaging-based diagnostics (e.g., histopathology, radiology), Grad-CAM demonstrates particular strength. For cause-of-death classification in colorectal cancer, LIME has provided valuable “explanations for both local and global level predictions.”¹¹¹ A comparative analysis specifically examining cancer patient data found that while both SHAP and LIME provided valuable insights,

the bagging algorithm had the best performance in all aspects when both XAI methods were applied.¹¹²

Factors driving XAI method selection in chronic disease applications. Several key considerations emerge as driving factors for XAI method selection across chronic disease types as follows:

(1) Data structure and modality

The nature of available clinical data strongly influences XAI method selection as follows:

- SHAP dominates for structured tabular clinical data.
- Grad-CAM is preferred for medical imaging applications.
- LIME shows versatility across data types but particularly excels with model-agnostic explanations.

(2) Explainability requirements

Different chronic diseases and clinical workflows have varying explainability needs, which include the following:

- When mathematical precision is paramount, SHAP tends to be preferred.
- When intuitive explanations for clinicians are prioritized, LIME often emerges as the choice.
- For visual explanations of image-based diagnostics, Grad-CAM is selected.

(3) Model compatibility

The underlying prediction models influence XAI method selection as follows:

- SHAP works particularly well with tree-based models like gradient-boosting decision trees.^{98,99}
- LIME offers model-agnostic capabilities valuable for diverse prediction approaches.¹¹⁰
- Grad-CAM is specifically designed for CNNs used in medical imaging.

The selection of specific XAI methods for chronic disease applications follows distinct patterns driven by data characteristics, explainability requirements, and the underlying prediction models. SHAP has emerged as dominant overall, particularly for structured clinical data analysis, while LIME offers advantages in accessibility and model-agnostic explanations. Grad-CAM maintains a specialized but crucial role in medical imaging applications.^{113–116}

As the field continues to evolve, researchers are increasingly employing multiple complementary XAI approaches to provide comprehensive explanations that address the

complex needs of chronic disease management across prediction, diagnosis, treatment planning, and long-term management. This review successfully quantified the distribution of XAI applications across these domains, with Figure 7 revealing that roughly 86% of the reviewed studies focused on disease prediction—underscoring the heavy emphasis on forecasting in chronic care. However, despite this progress, structural barriers continue to hinder the broader implementation of XAI in more complex domains such as the following:

Structural barriers to XAI implementation in treatment planning

(1) Decision complexity and validation challenges

Treatment planning involves significantly more complex decision-making processes than prediction tasks. While prediction models often have clear ground truth data for validation, treatment recommendations deal with multiple uncertainties and outcomes due to individual predispositions and genetic factors that make treatment validation considerably more challenging.^{117,118} The pluralistic nature of treatment decisions where multiple valid approaches may exist for the same condition further complicates the application of XAI in this domain.

(2) Clinician-AI interaction barriers

Survey evidence reveals significant usability challenges that limit XAI adoption in treatment planning. Radiotherapy planning studies show 88% of professionals report algorithm accuracy limitations, with 62% finding it difficult to modify AI-generated plans.¹¹⁷ These technical limitations create substantial barriers to integration in time-sensitive clinical workflows where treatment decisions must be made efficiently.

(3) Workflow integration challenges

Time constraints represent a major barrier to XAI implementation in treatment planning. Research on clinical conversations reveals that even traditional treatment planning discussions are frequently disrupted or deferred due to timing issues and competing clinical priorities.¹¹⁹ XAI systems that fail to account for these workflow realities face significant adoption barriers regardless of their technical capabilities.

Systemic challenges in XAI for disease management

(1) Multistakeholder coordination requirements

Chronic disease management inherently involves multiple stakeholders across care continuums, creating significant coordination challenges for XAI implementation. Research from Singapore's Primary Care Networks reveals

that effective chronic disease management requires systematic integration of ancillary services, monitoring systems, and funding streams level of coordination that current XAI systems struggle to address.¹²⁰

(2) Sociotechnical barriers

International research identifies four critical barrier categories to technology adoption in chronic disease management: Governance, technical, economic, and social factors.¹²¹ These barriers operate at multiple levels, from individual patient acceptance to institutional policy constraints, creating a complex implementation environment that purely technical XAI solutions cannot adequately address.

(3) Resource constraints and implementation disparities

Global research reveals significant disparities in technology implementation for chronic disease management, particularly in low- and middle-income countries. Clinical teams in resource-constrained environments see more patients with less time for evaluations and have reduced access to screening tools and diagnostic technologies.¹²² These resource limitations create substantial barriers to XAI implementation regardless of technical sophistication.

Technical challenges in advanced XAI applications

(1) Multimodal data integration complexities

Current XAI approaches show significant limitations in handling complex multimodal data necessary for comprehensive chronic disease management. While some innovative frameworks attempt to integrate tabular data, imaging, and genetic information using graph neural networks and region-based CNN approaches,¹²³ these systems remain limited in their ability to provide interpretable insights across heterogeneous data types.

(2) Intervention complexity modeling

Treatment interventions for chronic diseases often involve complex, multicomponent approaches with interactions that change over time. Current XAI systems struggle to model these dynamic intervention elements while maintaining explainability. Research on Alzheimer's treatment demonstrates attempts to address this through dashboard interfaces that visualize complex relationships, but significant challenges remain in representing intervention complexity.¹²³

(3) Real-world validation requirements

The transition from research to clinical implementation requires extensive real-world validation challenge

highlighted across multiple studies.^{124,118} Regulatory frameworks increasingly demand evidence of clinical utility and safety, while healthcare organizations require proof of integration with existing workflows. This validation gap represents a critical barrier to extending XAI beyond prediction into treatment and management domains.

Strategic recommendations for future research

(1) Building integrative XAI frameworks

Future research should focus on developing integrative XAI frameworks that bridge prediction, diagnosis, treatment planning, and management within unified systems. The example of combined knowledge graphs and CNN approaches for Alzheimer's treatment demonstrates potential pathways for creating more comprehensive XAI systems,¹²³ though significant work remains to extend these approaches across diverse chronic conditions.

(2) Addressing implementation science gaps

Incorporating implementation science perspectives that address the sociotechnical aspects of XAI adoption is essential for understanding how these systems can be effectively integrated into real-world healthcare settings, particularly by identifying the human, organizational, and technological factors that influence their uptake and sustained use. Studies examining how health information technology is implemented in chronic disease management reveal important facilitators and barriers that likely apply to XAI systems,¹²⁴ providing a conceptual framework for addressing implementation challenges.

Recent advances in LLM-based explainability for clinical decision support in chronic disease care. Recent advancements in large language models (LLMs) have introduced transformative capabilities for clinical decision support systems, particularly in XAI applications for chronic disease care. This expansion addresses critical gaps identified in traditional XAI approaches while introducing novel methods for enhancing transparency across the care continuum.

Emerging LLM-driven explainability paradigms

(1) Natural language explanation generation

Modern LLMs like GPT-4 demonstrate unprecedented ability to generate clinically coherent explanations through advanced natural language processing. Unlike traditional XAI methods that produce feature importance scores or activation maps, LLMs can contextualize model outputs within medical narratives. A comparative study of GPT-4 and human experts found 85% alignment in diagnostic reasoning patterns when

analyzing complex chronic disease cases.¹²⁵ This capability enables AI systems to mimic clinicians' cognitive processes while maintaining auditability—a critical advancement for chronic disease management requiring longitudinal reasoning. The integration of LLMs with established XAI techniques creates hybrid explanation systems. For instance, combining SHAP values with GPT-4's narrative generation produces both quantitative feature importance scores and clinically contextualized explanations.¹²⁴ This dual approach achieved 72% improvement in clinician comprehension scores compared to traditional XAI outputs in diabetes management trials.¹²⁶ However, despite these promising advances, LLMs have notable limitations. They are prone to hallucinations, generating plausible but factually incorrect information, which can mislead clinical decision-making.^{127–129} Additionally, noise and biases present in their large-scale training datasets may introduce inaccurate or outdated medical knowledge, undermining reliability and trustworthiness.^{127,130,131} These challenges necessitate careful validation and the development of robust safeguards when deploying LLM-based explanations in sensitive healthcare contexts.^{127,132–134}

(2) Multimodal clinical data interpretation

LLMs overcome traditional XAI limitations in handling heterogeneous medical data through advanced multimodal processing capabilities. A wound assessment system combining CNNs with GPT-4 demonstrated 89% accuracy in classifying chronic ulcer types while generating treatment recommendations aligned with vascular surgery guidelines.⁹⁰ This integration of visual and textual data processing represents a significant advancement over previous unimodal XAI systems. In chronic respiratory disease management, LLM-enhanced systems now process spirometry readings, CXRs, and patient-reported outcomes simultaneously. A recent implementation reduced diagnostic errors by 40% compared to single-modality approaches while providing unified explanations across data types.¹³⁵

LLM applications across care continuum

(1) Diagnostic decision support

GPT-4 shows particular promise in complex differential diagnosis for chronic conditions. When evaluating 110 clinical cases, GPT-4 achieved 78% diagnostic accuracy compared to 65% for traditional search-based methods.¹³⁶ The model's capacity to articulate diagnostic reasoning pathways in natural language addresses critical transparency requirements for chronic disease diagnosis, where multiple comorbidities often complicate clinical presentations. In neurological disorders, LLM-powered

systems now explain Alzheimer's progression predictions by correlating MRI findings with cognitive test results and medication histories. Clinicians report 30% faster diagnostic reconciliation using these systems compared to traditional XAI dashboards.¹²⁵

(2) Treatment personalization and planning

Augmenting LLMs with clinical practice guidelines creates powerful tools for chronic disease treatment optimization. A COVID-19 outpatient management system combining GPT-4 with WHO guidelines demonstrated 92% adherence to therapeutic protocols while explaining treatment choices in relation to individual patient characteristics.¹³⁷ This approach is now being adapted for chronic conditions like rheumatoid arthritis, where treatment plans require careful balancing of efficacy and side effect profiles. For bipolar depression management, guideline-augmented LLMs matched expert treatment recommendations in 51% of cases while avoiding contraindicated options 88% of the time.¹³⁸ The models' ability to explain pharmacological choices in relation to patient history and current symptoms represents a significant advance over previous black-box recommendation systems. In heart failure management, LLM-based systems reduced hospital readmissions by 22% through early identification of decompensation patterns and explainable intervention recommendations.¹³⁵ The integration of wearable device data with EHRs enabled continuous explanation generation aligned with individual patient trajectories.

Implementation challenges and considerations

(1) Explainability-accuracy tradeoffs

While LLMs enhance explanatory capabilities, they introduce new complexity in validation. A bipolar disorder study found that guideline-augmented models showed 23% lower accuracy than unmodified versions, highlighting the tension between clinical plausibility and predictive performance.¹³⁸ This emphasizes the need for specialized training protocols that balance explanatory depth with decision quality.

(2) Computational and ethical constraints

The resource intensity of LLM deployment poses significant barriers, particularly in resource-limited settings. A cost analysis revealed that implementing GPT-4-based chronic care systems requires 97% more computational resources than traditional XAI approaches.¹²⁶ However, techniques like retrieval-augmented generation and model distillation are reducing these requirements while maintaining explanatory fidelity.¹³⁵ Ethical concerns persist regarding potential automation bias in chronic care management.

Clinicians using LLM-powered systems showed 18% higher rates of uncritical acceptance compared to traditional XAI interfaces.¹³⁹ This underscores the need for careful interface design that maintains clinical oversight while leveraging LLM capabilities.

Strategic recommendations for LLM integration

1. Hybrid explanation systems: Combine LLMs with traditional XAI techniques to leverage both quantitative feature attribution and narrative explanations.^{124,135}
2. Guideline-compliant architectures: Implement constrained generation frameworks that tether LLM outputs to established clinical protocols.^{137,140}
3. Multimodal validation protocols: Develop testing frameworks that assess both explanatory quality and clinical accuracy across data types.^{90,135}
4. Resource-optimized deployment: Utilize techniques like prompt engineering and model distillation to enable LLM implementation across healthcare settings.^{126,141}

In summary, the integration of LLMs into chronic disease XAI systems represents a paradigm shift in clinical decision support. By bridging the explanatory gap between complex AI outputs and clinical reasoning processes, these advanced models address critical limitations of traditional XAI approaches. However, successful implementation requires careful attention to validation protocols, computational constraints, and human-AI collaboration dynamics. Future research should focus on developing standardized evaluation metrics for LLM-based explanations while ensuring these systems complement rather than replace clinician expertise in chronic care management.

Limitations and future research

This systematic review has achieved several significant objectives, providing a comprehensive understanding of XAI applications in chronic disease healthcare. It effectively mapped the relationship between specific XAI algorithms and various chronic conditions, demonstrating how different approaches like SHAP, LIME, and Grad-CAM align with particular medical data types and diagnostic needs. The review has also provided valuable insights into the effectiveness of different XAI approaches across various chronic conditions, offering practical guidance for future implementations. However, several important limitations should be acknowledged when interpreting the review's findings.

Fundamental limitations of current XAI approaches

As Ghassemi et al. discussed, current approaches to explainability, including widely used methods like SHAP

and LIME, offer a “false hope” that users can effectively judge AI decisions based on local explanations.¹⁴² In fact, these explanation methods fundamentally rely on humans to interpret what the explanations mean, often leading to confirmation bias, where people assume the model uses the same features they would find important. These methods suffer from the same interpretability problems as other techniques like saliency mapping. LIME attempts to understand decisions by altering inputs and observing changes in outputs, but still leaves the interpretation of these results to humans. Perhaps most problematically, evidence suggests that explainability techniques may actually decrease vigilance among users. Models with explanations can hamper people’s ability to detect serious mistakes and may unreasonably increase confidence in algorithmic decisions.

As alternatives to current explainability requirements, the authors advocate for treating AI systems as true black boxes and focusing on rigorous validation across diverse populations. They also recommend using RCTs as the gold standard for evaluating AI interventions, just as we do with other medical technologies.¹⁴²

Moreover, as Rudin discussed, a fundamental issue is that explanations for black-box models are inherently unreliable.¹⁴³ Since these explanations cannot perfectly represent what the original model computes (otherwise they would be the model itself), explanations inevitably contain inaccuracies. When an explanation model is wrong 10% of the time, healthcare providers cannot trust either the explanation or the underlying black box, potentially leading to medical errors and patient harm.

Clinical implementation challenges

Healthcare decision-making is inherently complex and requires integrating information from multiple sources. Black box models create significant challenges when clinicians need to combine model predictions with factors outside the database. For example, if a risk prediction model doesn’t account for a critical patient characteristic, physicians have no clear way to calibrate how much this additional information should adjust the estimated risk. Transparent models allow practitioners to see exactly what factors are being considered and make appropriate clinical judgments.

Implementation complexity is another practical concern in healthcare settings. Complex models requiring numerous inputs are prone to data entry errors. These errors have been documented to affect treatment decisions, creating a form of procedural unfairness where identical patients might receive different care due to random input mistakes. The medical field is particularly vulnerable to confounding variables and data biases. Interpretable models make these issues more detectable and addressable. With transparent models, such critical errors would be immediately apparent, preventing potentially dangerous diagnostic mistakes. It is

suggested that inherently interpretable models be designed from the start rather than creating complicated black box models and then trying to explain them afterward. This approach avoids the fundamental problems with explanations that are never fully faithful to the original model. Also, it is suggested that one possible policy mandate could require that “no black box should be deployed when there exists an interpretable model with the same level of performance.”¹⁴³

Evolving clinical perspectives on XAI

On the other hand, as Giacobbe and Bassetti discussed, while past studies emphasized explanations of predictive factors that clinicians could easily incorporate into their clinical reasoning, modern studies using AI are increasingly focusing on performance metrics instead.¹⁴⁴ This shift creates several challenges: Current XAI methods have reliability limitations, performance metrics don’t translate easily to clinical decision-making, and the reduced prominence of explanations may make AI models seem unfathomable to clinicians, potentially increasing skepticism toward AI-based tools.

While acknowledging that current explanations offered by techniques such as SHAP and LIME are not flawless, the authors caution against their blacklisting from medical literature. They recommend promoting clinicians’ understanding of these structural changes and emphasize that the goal should be to reduce rather than increase skepticism toward medical AI through better communication.

Limitations of methodology

The temporal scope of the analysis presents a significant limitation, as the rapid evolution of XAI algorithms and their applications means that more recent developments may not be captured in the reviewed literature. This particularly affects our understanding of the latest advancements in algorithm performance and real-world implementations.

Data limitations manifest in multiple ways throughout this review. Many analyzed studies reported insufficient data volume, limited patient history horizons, or class imbalance issues. These underlying data constraints affect our ability to draw robust conclusions about the effectiveness of different XAI approaches. Additionally, our reliance on major academic databases may have excluded relevant research published in regional, specialized, or emerging databases, potentially missing valuable insights from less prominently indexed sources, particularly from developing countries or specialized medical institutions.

This review faces substantial limitations in assessing real-world clinical utility, as many studies focused on theoretical implementations or controlled research settings rather than actual clinical environments. This gap between research findings and practical implementation makes it

challenging to fully evaluate the true potential of XAI in chronic disease care.

Our ability to compare effectiveness across different XAI approaches is limited by the heterogeneity in how different studies evaluated and reported their results. This variation in reporting metrics and evaluation methods makes it difficult to conduct direct comparisons between different XAI implementations.

Publication and selection bias

Publication bias represents another significant limitation that must be carefully considered. Studies showing positive or significant results regarding XAI applications in chronic diseases may have been more likely to be published than those showing negative or inconclusive results. This potential bias could lead to an overestimation of the effectiveness of certain XAI algorithms in healthcare applications. Furthermore, the review may not capture unpublished work, ongoing studies, or proprietary research being conducted within healthcare institutions and companies. This limitation particularly affects our understanding of XAI applications in real-world clinical settings, as unsuccessful implementation attempts or challenges may be underreported in the published literature. Language restrictions in publication may have also contributed to this bias, potentially excluding valuable insights from non-English language studies.

Future research directions

Future research initiatives should address these limitations through several key strategies. These include implementing more rigorous and comprehensive data collection methodologies, expanding database coverage across diverse geographical and cultural contexts, and developing standardized evaluation metrics that facilitate meaningful cross-study comparisons. Additionally, future studies should actively seek to document and analyze both successful and unsuccessful implementation attempts, along with their associated challenges, to provide a more balanced representation in the published literature. Moreover, while this review aimed to provide a comprehensive overview of XAI algorithms applied across a wide range of chronic diseases, the breadth of topics covered inevitably limited the depth of discussion for each individual condition. As a result, some disease-specific insights may remain somewhat general or high-level. This broad scope was intentional, as it allows for the identification of overarching patterns and cross-cutting challenges in the field. However, future research would benefit from more focused systematic reviews that delve deeper into the use of XAI in the context of specific chronic diseases or more narrowly defined clinical applications. Such dedicated analyses could provide more nuanced insights and practical

recommendations tailored to particular patient populations and healthcare needs.

Of particular importance is the need to expand research beyond purely technical aspects of XAI implementations to examine their broader societal impacts, including effects on human decision-making capabilities and health outcomes.¹⁴⁵ As demonstrated by Noorbehbahani et al.,¹⁴⁶ digital technologies can create interconnected patterns of psychological challenges that affect significant portions of the population. Similarly, the implementation of XAI systems in healthcare contexts may produce complex effects on provider workflows, patient-provider relationships, and treatment adherence that extend well beyond the immediate diagnostic or predictive accuracy of the models. This multifaceted approach would contribute to a more nuanced and complete understanding of XAI applications in healthcare settings.

Conclusion

Based on our comprehensive systematic review, several key insights emerge regarding the application of XAI in chronic disease healthcare. A notable finding is the significant imbalance in XAI applications, with 86.2% of studies focusing on disease prediction rather than management, diagnosis, or treatment. This concentration reveals both current research priorities and critical opportunities for expansion.

The prevalence of certain XAI algorithms, particularly SHAP, LIME, and Grad-CAM, emerges as a significant pattern in the literature. SHAP's dominance can be attributed to its mathematical rigor through game theory foundations and its versatility with structured clinical data. LIME's popularity stems from its ability to provide locally interpretable explanations crucial for individual patient cases. Grad-CAM's prominence in specific diseases, especially those requiring medical imaging, demonstrates how the choice of XAI algorithm often aligns with the nature of the medical data being analyzed.

The review highlights interesting patterns in disease coverage. Chronic conditions with clear diagnostic criteria and structured data, such as CKD and diabetes, receive more attention in XAI applications. In contrast, conditions requiring more complex, multimodal data or those with less standardized diagnostic processes show limited XAI implementation. This disparity reflects both the technical challenges in applying XAI to complex medical data and the varying availability of structured healthcare datasets across different conditions.

To address the predominant focus on prediction, several promising approaches for expanding XAI applications into treatment and management domains emerge from our analysis. These include developing temporal XAI frameworks to track treatment effectiveness over time, integrating XAI into clinical decision support systems for treatment monitoring, and creating interactive interfaces for patient education and engagement. Additionally, XAI can be adapted to

support personalized treatment planning by analyzing historical treatment outcomes and explaining treatment recommendations based on individual patient characteristics. These applications could be particularly valuable for optimizing resource allocation and improving the efficiency of chronic disease management programs.

Looking forward, we recommend several key areas for advancement: (1) Development of specialized XAI algorithms that can handle the temporal nature of chronic disease progression, (2) integration of domain knowledge into XAI frameworks to produce more clinically relevant explanations, (3) creation of standardized benchmarking datasets for evaluating healthcare XAI systems, (4) investment in prospective studies to validate the clinical utility of XAI-enabled systems in real-world healthcare settings, and (5) Development of standardized evaluation frameworks to assess the effectiveness of expanded XAI applications in improving treatment and management outcomes.

The successful integration of XAI in chronic disease care will require close collaboration between AI researchers, healthcare professionals, and domain experts to ensure that these systems not only provide accurate predictions but also deliver actionable insights that can improve patient outcomes. As the field matures, focusing on expanding XAI applications beyond prediction while maintaining interpretability and clinical relevance will be crucial for advancing the practical implementation of XAI in healthcare settings. This expansion, coupled with rigorous validation and standardization efforts, will help realize the full potential of XAI in transforming chronic disease care.

Acknowledgments

We would like to express our sincere gratitude to the members of the Persuasive Design Lab at York University for their valuable insights and constructive feedback throughout the writing of this paper.

ORCID iD

Kiemute Oyibo  <https://orcid.org/0000-0001-8300-3343>

Author contributions

Hooman Hoghooghi Esfahani: Data curation, formal analysis, investigation, methodology, software, visualization, writing—original draft. **Shogo Toyonaga:** Data curation, project administration, resources, writing—review & editing. **Kiemute Oyibo:** Conceptualization, funding acquisition, project administration, supervision, validation, writing—review & editing.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was funded by the corresponding author's Natural Sciences and Engineering Research Council of Canada

(NSERC) Discovery Grant (RGPIN–2023–0519) and the Connected Minds research program at York University.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement

All data generated or analyzed during this review are included in this published article and its supplemental information files.

Supplemental material

Supplemental material for this article is available online.

References

1. Goodman RA, Posner SF, Huang ES, et al. Defining and measuring chronic conditions: imperatives for research, policy, program, and practice. *Prev Chronic Dis* 2013; 10: E66.
2. Backe MB, Kallestrup P, Rasmussen K, et al. Burden of selected chronic non-communicable diseases in a primary healthcare setting in nuuk, Greenland, compared to a danish suburb. *Scand J Prim Health Care* 2024; 42: 435–441.
3. Chan SWC. Chronic disease management, self-efficacy and quality of life. *J Nurs Res* 2021; 29: e129.
4. Kuru Alici N and Öztürk Çopur E. Nurses' experiences as care providers for Syrian refugees with noncommunicable diseases: a qualitative study. *J Transcult Nurs* 2023; 34: 24–31.
5. Longhurst G. Community revitalisation strategy. In 2017. Available from: <https://api.semanticscholar.org/CorpusID:80328960>.
6. Wilson PWF, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; 97: 1837–1847.
7. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 2020; 58: 82–115.
8. Holzinger A, Langs G, Denk H, et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; 9: e1312.
9. Tjoa E and Guan C. A survey on explainable artificial intelligence (xai): toward medical xai. *IEEE Trans Neural Networks Learn Syst* 2020; 32: 4793–4813.
10. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020; 20: 1–9.
11. Tonekaboni S, Joshi S, McCradden MD, et al. What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Machine learning for healthcare conference. PMLR*, 2019. pp. 359–80.
12. Ribeiro MT, Singh S and Guestrin C. “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, 2016. pp. 1135–44.
13. Lundberg S. A unified approach to interpreting model predictions. *arXiv Prepr arXiv170507874*. 2017.
 14. Wachter S, Mittelstadt B and Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv JL Tech* 2017; 31: 841.
 15. Vaswani A. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30: 1–11.
 16. Kumar M. AI-Powered Wearables for Continuous Patient Monitoring: Improving Chronic Disease Management and Preventive Care. *Int Sci J Eng Manag [Internet]* 2024; 1: 1–8. Available from: <https://api.semanticscholar.org/CorpusID:274613516>.
 17. Ulgu MM, Erturkmen GBL, Yuksel M, et al. A nationwide chronic disease management solution via clinical decision support services: software development and real-life implementation report. *JMIR Med Informatics* 2024; 12: e49986.
 18. Duda-Sikuła M and Kurpas D. Enhancing chronic disease management: personalized medicine insights from rural and urban general practitioner practices. *J Pers Med* 2024; 14: 706.
 19. Law B, Chhatwal PK, Licskai C, et al. Patient engagement in interprofessional team-based chronic disease management: a qualitative description of a Canadian program. *Patient Educ Couns* 2023; 114: 107836.
 20. Burgermaster M, Rosenthal M, Tierney WM, et al. Nutri: a behavioral science-based clinical decision support for chronic disease management. In: *AMIA Annual Symposium Proceedings*, 2023. p. 299.
 21. Li S. Effects of clinical decision support systems in chronic disease management. *Int J Clin Exp Med [Internet]* 2024; 17: 1–11. Available from: <https://api.semanticscholar.org/CorpusID:269542936>.
 22. Aziz NA, Manzoor A, Mazhar Qureshi MD, et al. Explainable AI in Healthcare: Systematic Review of Clinical Decision Support Systems. *medRxiv*. 2024;2008–24.
 23. Guleria P, Naga Srinivasu P, Ahmed S, et al. XAI Framework for cardiovascular disease prediction using classification techniques. *Electronics (Basel)* 2022; 11: 4086.
 24. Sethi A, Dharmavaram S and Somasundaram SK. Explainable Artificial Intelligence (XAI) Approach to Heart Disease Prediction. In: *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*. IEEE, 2024. pp. 1–6.
 25. JAGADEESH P. Development and evaluation of an explainable ai model for early chronic kidney disease diagnosis. *Int J Mech Eng Res Technol* 2024; 16: 77–92.
 26. Moreno-Sánchez PA. Data-driven early diagnosis of chronic kidney disease: development and evaluation of an explainable AI model. *IEEE Access* 2023; 11: 38359–38369.
 27. Chandler J, Cumpston M, Li T, et al. *Cochrane handbook for systematic reviews of interventions*. Hoboken: Wiley, 2019.
 28. Cumpston M, Li T, Page MJ, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. *Cochrane database Syst Rev* 2019; 2019: 1–2.
 29. Group GW. Grading quality of evidence and strength of recommendations. *Br Med J* 2004; 328: 1490.
 30. Schünemann HJ, Higgins JPT, Vist GE, et al. Completing ‘summary of findings’ tables and grading the certainty of the evidence. *Cochrane Handb Syst Rev Interv* 2019: 375–402.
 31. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Br Med J* 2021; 372: 1–9.
 32. Landis JR. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 1977; 33: 159–174.
 33. Rodriguez Nava G, Miranti E, McIntyre K, et al. A machine learning exploration of social determinants of health and hospital-onset bacteremia, northern California, 2019–2023. *Antimicrob Steward Healthc Epidemiol [Internet]*. 2024/09/16, 2024; 4: s132–s133. Available from: <https://www.cambridge.org/core/product/A61FB5ED055DCB0CC3DFAA4938DF6D08>.
 34. Chawki Y, Elasnoui K and Ouhda M. Classification and detection of COVID-19 based on X-ray and CT images using deep learning and machine learning techniques: a bibliometric analysis. *AIMS Electron Electr Eng* 2024; 8: 71–103.
 35. Barnett T, Tollit M, Ratnapalan S, et al. Education support services for improving school engagement and academic performance of children and adolescents with a chronic health condition. *Cochrane Database Syst Rev* 2023: 1–58.
 36. Kamal MS, Chowdhury L, Nimmy SF, et al. An Interpretable Framework for Identifying Cerebral Microbleeds and Alzheimer’s Disease Severity using Multimodal Data. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023. pp. 1–4.
 37. Sekaran K, Alsamman AM, George Priya Doss C, et al. Bioinformatics investigation on blood-based gene expressions of Alzheimer’s disease revealed ORAI2 gene biomarker susceptibility: an explainable artificial intelligence-based approach. *Metab Brain Dis* 2023; 38: 1297–1310.
 38. Prajapati J, Uduthalapally V, Das D, et al. XAIA: An Explainable AI Approach for Classification and Analysis of Blood Anemia. In: *2023 OITS International Conference on Information Technology (OCIT)*. IEEE, 2023. pp. 88–93.
 39. Narteni S, Baiardini I, Braidò F, et al. Explainable artificial intelligence for cough-related quality of life impairment prediction in asthmatic patients. *PLoS One* 2024; 19: e0292980.
 40. Maouche I, Terrissa LS, Benmohammed K, et al. An explainable ai approach for breast cancer metastasis prediction based on clinicopathological data. *IEEE Trans Biomed Eng* 2023; 70: 3321–3329.

41. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020; 2: 56–67.
42. Jhumka K, Auzine MM, Heenaye-Mamode Khan M, et al. Explainable Chronic Kidney Disease (CKD) Prediction using Deep Learning and Shapley Additive Explanations (SHAP). In: *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence*, 2023. pp. 29–33.
43. Manju VN and Aparna N. Decision Tree-Based Explainable AI for Diagnosis of Chronic Kidney Disease. In: *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2023. pp. 947–52.
44. Vijayvargiya A, Raghav A, Bhardwaj A, et al. A lime-based explainable machine learning technique for the risk prediction of chronic kidney disease. In: *2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3)*. IEEE, 2023. pp. 1–6.
45. Islam MA, Nittala K and Bajwa G. Adding explainability to machine learning models to detect chronic kidney disease. In: *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2022. pp. 297–302.
46. Mehedi MHK, Haque E, Radin SY, et al. Kidney tumor segmentation and classification using deep neural network on ct images. In: *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2022. pp. 1–7.
47. Jhumka K, Khan MHM, Mungloo-Dilmohamud Z, et al. Classification of kidney abnormalities using deep learning with explainable AI. In: *2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*. IEEE, 2023. pp. 133–7.
48. Shang Y, Tian Y, Zhou M, et al. EHR-oriented knowledge graph system: toward efficient utilization of non-used information buried in routine clinical practice. *IEEE J Biomed Heal Informatics* 2021; 25: 2463–2475.
49. Paul S and Al Mamun M. An Explainable Machine Learning Model for Diagnosis of Chronic Kidney Disease Using Regular Health Informatics. In: *2023 5th International Conference on Sustainable Technologies for Industry 50 (STI)*. IEEE, 2023. pp. 1–6.
50. Vásquez-Morales GR, Martínez-Monterrubio SM, Moreno-Ger P, et al. Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning. *Ieee Access* 2019; 7: 152900–10.
51. Ghosh SK and Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci Rep* 2024; 14: 3687.
52. Hamed O, Soliman A and Etmnani K. Temporal Context Matters: An Explainable Model for Medical Resource Utilization in Chronic Kidney Disease. In: *MIE*, 2023. pp. 613–4.
53. Arumugham V, Sankaralingam BP, Jayachandran UM, et al. An explainable deep learning model for prediction of early-stage chronic kidney disease. *Comput Intell* 2023; 39: 1022–1038.
54. Moreno-Sanchez PA. An automated feature selection and classification pipeline to improve explainability of clinical prediction models. In: *2021 IEEE 9th international conference on healthcare informatics (ICHI)*. IEEE, 2021. pp. 527–34.
55. Ikechukwu AV and Murali S. 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)*. IEEE, 2023. pp. 1–6.
56. Ikechukwu AV, Murali S and Honnaraju B. COPDNet: An Explainable ResNet50 Model for the Diagnosis of COPD from CXR Images. In: *2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON)*. IEEE, 2023. pp. 1–7.
57. Wang X, Qiao Y, Cui Y, et al. An explainable artificial intelligence framework for risk prediction of COPD in smokers. *BMC Public Health* 2023; 23: 2164.
58. Vaccari I, Orani V, Paglialonga A, et al. A generative adversarial network (gan) technique for internet of medical things data. *Sensors* 2021; 21: 3726.
59. El-Magd LMA, Dahy G, Farrag TA, et al. An interpretable deep learning based approach for chronic obstructive pulmonary disease using explainable artificial intelligence. *Int J Inf Technol* 2024; 16: 1–16.
60. Wu Y, Li X, Zhang X, et al. Community-Based Hierarchical Positive-Unlabeled (PU) Model Fusion for Chronic Disease Prediction. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023. pp. 2747–56.
61. Ahmad A and Mehruz S. Explainable AI-based early detection of diabetes for smart healthcare. 2022.
62. Shaheen I, Javaid N, Alrajeh N, et al. Hi-Le and HiTCL: Ensemble Learning Approaches for Early Diabetes Detection using Deep Learning and eXplainable Artificial Intelligence. *IEEE Access* 2024; 12: 1–23.
63. Nagaraj P, Muneeswaran V, Dharanidharan A, et al. A prediction and recommendation system for diabetes mellitus using XAI-based lime explainer. In: *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 2022. pp. 1472–8.
64. Kibria HB, Nahiduzzaman M, Goni MOF, et al. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors* 2022; 22: 7268.
65. Joseph LP, Joseph EA and Prasad R. Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. *Comput Biol Med* 2022; 151: 106178.
66. Ganguly R and Singh D. Explainable artificial intelligence (xai) for the prediction of diabetes management: An ensemble approach. *Int J Adv Comput Sci Appl* 2023; 14: 1–6.
67. Moreno-Sánchez PA, Arroyo-Fernández R, Bravo-Esteban E, et al. Assessing the relevance of mental health factors in fibromyalgia severity: a data-driven case study using explainable AI. *Int J Med Inform* 2024; 181: 105280.

68. Huang M, Zhang XS, Bhatti UA, et al. An interpretable approach using hybrid graph networks and explainable AI for intelligent diagnosis recommendations in chronic disease care. *Biomed Signal Process Control* 2024; 91: 105913.
69. Pfeuffer N. Design Principles for (X) AI-based Patient Education Systems. 2021.
70. Rodríguez-Belenguer P, Piñana JL, Sánchez-Montañés M, et al. A machine learning approach to identify groups of patients with hematological malignant disorders. *Comput Methods Programs Biomed* 2024; 246: 108011.
71. Deo R and Panigrahi S. Explainability analysis of black box SVM models for hepatic steatosis screening. In: *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*. IEEE, 2022. pp. 22–5.
72. Ahmed R and Imran AS. Knee Osteoarthritis Analysis Using Deep Learning and XAI on X-rays. *IEEE Access* 2024; 12: 1–10.
73. Baweja AK, Aditya S and Kanchana M. Leprosy Diagnosis using Explainable Artificial Intelligence Techniques. In: *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 2023. pp. 551–6.
74. Arya G, Bagwari A, Saini H, et al. Explainable AI for enhanced interpretation of liver cirrhosis biomarkers. *IEEE Access* 2023; 11: 1–13.
75. Sharma R, Mangla M, Patil S, et al. Lung Disease Detection from Chest X-Ray Using GANs. In: *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 2024. pp. 565–72.
76. Koyyada SP and Singh TP. An AI Decision System to Predict Lung Nodules through Localization from Chest X-ray Images. In: *2023 9th International Conference on Signal Processing and Communication (ICSC)*. IEEE, 2023. pp. 214–20.
77. Choi Y and Lee H. Interpretation of lung disease classification with light attention connected module. *Biomed Signal Process Control* 2023; 84: 104695.
78. Morabito F, Adornetto C, Monti P, et al. Genes selection using deep learning and explainable artificial intelligence for chronic lymphocytic leukemia predicting the need and time to therapy. *Front Oncol* 2023; 13: 1198992.
79. Pezoulas VC, Kalatzis F, Exarchos TP, et al. A federated AI-empowered platform for disease management across a Pan-European data driven hub. In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2022. pp. 1–4.
80. De Faria TP, Do Nascimento MZ and Martins LGA. Understanding the multiclass classification of lymphomas from simple descriptors. In: *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2021. pp. 1202–8.
81. Benmohammed K, Valensi P, Omri N, et al. Metabolic syndrome screening in adolescents: new scores AI_METS based on artificial intelligence techniques. *Nutr Metab Cardiovasc Dis* 2022; 32: 2890–2899.
82. Yagin FH, Shateri A, Nasiri H, et al. Development of an expert system for the classification of myalgic encephalomyelitis/chronic fatigue syndrome. *PeerJ Comput Sci* 2024; 10: e1857.
83. Yagin FH, Alkhateeb A, Raza A, et al. An explainable artificial intelligence model proposed for the prediction of myalgic encephalomyelitis/chronic fatigue syndrome and the identification of distinctive metabolites. *Diagnostics* 2023; 13: 3495.
84. Davagdorj K, Bae JW, Pham VH, et al. Explainable artificial intelligence based framework for non-communicable diseases prediction. *Ieee Access* 2021; 9: 123672–88.
85. Brakefield WS, Ammar N and Shaban-Nejad A. An urban population health observatory for disease causal pathway analysis and decision support: underlying explainable artificial intelligence model. *JMIR Form Res* 2022; 6: e36055.
86. Troncoso-García AR, Martínez-Ballesteros M, Martínez-Álvarez F, et al. Explainable machine learning for sleep apnea prediction. *Procedia Comput Sci* 2022; 207: 2930–2939.
87. Junaid M, Ali S, Eid F, et al. Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease. *Comput Methods Programs Biomed* 2023; 234: 107495.
88. Ukwuoma CC, Qin Z, Heyat MB, et al. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *J Adv Res* 2023; 48: 191–211.
89. Marvin G and Alam MGR. Explainable augmented intelligence and deep transfer learning for pediatric pulmonary health evaluation. In: *2022 international conference on innovations in science, engineering and technology (ICISSET)*. IEEE, 2022. pp. 272–7.
90. Lo ZJ, Mak MHW, Liang S, et al. Development of an explainable artificial intelligence model for Asian vascular wound images. *Int Wound J* 2024; 21: e14565.
91. Uegami W, Bychkov A, Ozasa M, et al. MIXTURE Of human expertise and deep learning—developing an explainable model for predicting pathological diagnosis and survival in patients with interstitial lung disease. *Mod Pathol* 2022; 35: 1083–1091.
92. Li J, Wang C, Chen J, et al. Explainable CNN with fuzzy tree regularization for respiratory sound analysis. *IEEE Trans Fuzzy Syst* 2022; 30: 1516–1528.
93. McGuinness LA and Higgins JPT. Risk-of-bias VISualization (robvis): an R package and shiny web app for visualizing risk-of-bias assessments. *Res Synth Methods* 2021; 12: 55–61.
94. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*, 2017. pp. 618–26.
95. Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol* 2017; 10: 257–273.

96. Daanouni O, Cherradi B and Tmiri A. Automatic detection of diabetic retinopathy using custom cnn and grad-cam. In: *Advances on Smart and Soft Computing: Proceedings of ICACIn 2020*. Springer, 2021. pp. 15–26.
97. Loecher M. Debiasing SHAP scores in random forests. *ASIA Adv Stat Anal* 2024; 108: 427–440.
98. Baniecki H and Biecek P. Manipulating shap via adversarial data perturbations (student abstract). In: *Proceedings of the aaai conference on artificial intelligence*, 2022. pp. 12907–8.
99. Drapala J and Świątek J. Generating synthetic mixed-type tabular data by decoding samples from a latent-space: a case study in healthcare. *Procedia Comput Sci* 2024; 246: 2254–2263.
100. Wang Z, Gao C, Xiao C, et al. MediTab: scaling medical tabular data predictors via data consolidation, enrichment, and refinement. *arXiv Prepr arXiv230512081*. 2023.
101. Abhadiomhen SE, Nzeakor EO and Oyibo K. Health risk assessment using machine learning: systematic review. *Electronics (Basel)* 2024; 13: 4405.
102. Okolo CT. Optimizing human-centered AI for healthcare in the Global South. *Patterns* 2022; 3: 1–7.
103. Kasaye MD, Getahun AG and Kalayou MH. Exploring artificial intelligence for healthcare from the health professionals' perspective: the case of limited resource settings. *Digit Heal* 2025; 11: 20552076251330550.
104. Okolo CT. AI in the “real world”: examining the impact of AI deployment in low-resource contexts. *arXiv Prepr arXiv201201165*. 2020.
105. Vimbi V, Shaffi N and 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.
106. Duvvur V. Enhancing Explainability in AI Models: A Quantitative Comparison of XAI Techniques for Large Language Models and Healthcare Applications. *J Inf Syst Eng Manag [Internet]* 2025; 10: 1–51. Available from: <https://api.semanticscholar.org/CorpusID:277800419>.
107. Wang Y. A comparative analysis of model agnostic techniques for explainable artificial intelligence. *Res Reports Comput Sci* 2024; 3: 25–33.
108. Balve AK and Hendrix P. Interpretable breast cancer classification using CNNs on mammographic images. *arXiv Prepr arXiv240813154*. 2024.
109. Hooshyar D and Yang Y. Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access* 2024; 12: 1–19.
110. Nayeibi A, Tipirneni S, Foreman B, et al. An empirical comparison of explainable artificial intelligence methods for clinical data: a case study on traumatic brain injury. In: *AMIA annual symposium proceedings*, 2023. pp. 815.
111. Lee SY, Chu WCC, Tseng YH, et al. Explainable AI Applied in Healthcare: A Case Study of Diabetes Prediction. In: *2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. IEEE, 2024. pp. 336–44.
112. Lázaro C and Angulo C. Iterative Application of UMAP-Based Algorithms for Fully Synthetic Healthcare Tabular Data Generation. 2024.
113. Ennab M and Mcheick H. Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and grad-CAM models. *Mach Learn Knowl Extr* 2025; 7: 12.
114. Mahesh TR, Vinoth Kumar V and Guluwadi S. Enhancing brain tumor detection in MRI images through explainable AI using grad-CAM with resnet 50. *BMC Med Imaging* 2024; 24: 107.
115. Yiğit T, Şengöz N, Özmen Ö, et al. Diagnosis of paratuberculosis in histopathological images based on explainable artificial intelligence and deep learning. *arXiv Prepr arXiv220801674*. 2022.
116. Suara S, Jha A, Sinha P, et al. Is grad-cam explainable in medical images? In: *International Conference on Computer Vision and Image Processing*. Springer, 2023. pp. 124–35.
117. Sung M, He J, Zhou Q, et al. Using an integrated framework to investigate the facilitators and barriers of health information technology implementation in noncommunicable disease management: systematic review. *J Med Internet Res* 2022; 24: e37338.
118. Wurst C, Winnie Chen HY, Sage M, et al. Critical decision method interviews to understand the initial treatment planning process in foster care. In: *Proceedings of the human factors and ergonomics society annual meeting*. Los Angeles, CA: Sage Publications Sage CA, 2022, pp.1060–1064.
119. De Foo C, Surendran S, Tam CH, et al. Perceived facilitators and barriers to chronic disease management in primary care networks of Singapore: a qualitative study. *BMJ Open* 2021; 11: e046010.
120. Kamal MS, Dey N, Chowdhury L, et al. Explainable AI for glaucoma prediction analysis to understand risk factors in treatment planning. *IEEE Trans Instrum Meas* 2022; 71: 1–9.
121. Petragallo R, Bardach N, Ramirez E, et al. Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: a survey study of medical dosimetrists. *J Appl Clin Med Phys* 2022; 23: e13568.
122. Parvin S, Nimmy SF and Kamal MS. Convolutional neural network based data interpretable framework for Alzheimer's treatment planning. *Vis Comput Ind Biomed Art* 2024; 7: 3.
123. Dadkhah M, Mehraeen M, Rahimnia F, et al. Exploring the experts' perceptions of barriers to using internet of things for chronic disease management in Iran. *J Sci Technol Policy Manag* 2023; 14: 440–458.
124. Mohamed YA, Khoo BE, Asaari MSM, et al. Decoding the black box: explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning-A state-of-the art systematic review. *Int J Med Inform* 2024; 191: 105689.
125. Hou J and Wang LL. Explainable AI for Clinical Outcome Prediction: A Survey of Clinician Perceptions and Preferences. *arXiv Prepr arXiv250220478*. 2025.

126. Kashyap AM, Rao D, Boland MR, et al. Predicting explainable dementia types with LLM-aided feature engineering. *Bioinformatics* 2025; 41: btaf156.
127. Tonmoy SM, Zaman SM, Jain V, et al. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv Prepr arXiv240101313*. 2024;6.
128. Zhao X, Yu J, Liu Z, et al. Medico: Towards Hallucination Detection and Correction with Multi-source Evidence Fusion. *arXiv Prepr arXiv241010408*. 2024.
129. Pandit S, Xu J, Hong J, et al. MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models. *arXiv Prepr arXiv250214302*. 2025.
130. Lammert J, Dreyer TF, Lörsch AM, et al. *Large language models for precision oncology: Clinical decision support through expert-guided learning*. Vol. 42. Alexandria, VA: American Society of Clinical Oncology, 2024.
131. Poulain R, Fayyaz H and Beheshti R. Bias patterns in the application of llms for clinical decision support: a comprehensive study. *arXiv Prepr arXiv240415149*. 2024.
132. Anjum S, Zhang H, Zhou W, et al. HALO: Hallucination Analysis and Learning Optimization to Empower LLMs with Retrieval-Augmented Context for Guided Clinical Decision Making. *arXiv Prepr arXiv240910011*. 2024.
133. Ahmad MA, Yaramis I and Roy TD. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv Prepr arXiv231101463*. 2023.
134. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024; 30: 2613–2622.
135. Al Machot F, Horsch MT and Ullah H. Building Trustworthy AI: Transparent AI Systems via Large Language Models, Ontologies, and Logical Reasoning (TranspNet). *arXiv Prepr arXiv241108469*. 2024.
136. Sandmann S, Riepenhausen S, Plagwitz L, et al. Systematic analysis of ChatGPT, google search and llama 2 for clinical decision support tasks. *Nat Commun* 2024; 15: 2050.
137. Oniani D, Wu X, Visweswaran S, et al. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In: *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2024. pp. 694–702.
138. Perlis RH, Goldberg JF, Ostacher MJ, et al. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology* 2024; 49: 1412–1416.
139. Elbattah M, Arnaud E, Ghazali DA, et al. Exploring the Ethical Challenges of Large Language Models in Emergency Medicine: A Comparative International Review. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024. pp. 5750–5.
140. Mumuni F and Mumuni A. Explainable artificial intelligence (XAI): from inherent explainability to large language models. *arXiv Prepr arXiv250109967*. 2025.
141. Nazary F, Deldjoo Y and Di Noia T. ChatGPT-HealthPrompt. Harnessing the power of XAI in prompt-based healthcare decision support using ChatGPT. In: *European Conference on Artificial Intelligence*. Springer, 2023. pp. 382–97.
142. Ghassemi M, Oakden-Rayner L and Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Heal* 2021; 3: e745–e750.
143. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; 1: 206–215.
144. Giacobbe DR and Bassetti M. The fading structural prominence of explanations in clinical studies. *Int J Med Inform* 2025; 197: 105835.
145. Noorbehbahani F, Zaremohzzabieh Z, Esfahani HH, et al. AI on Loss in Decision-Making and Its Associations With Digital Disorder, Socio-Demographics, and Physical Health Outcomes in Iran. In: *Exploring Youth Studies in the Age of AI. IGI Global*, 2024. pp. 254–65.
146. Noorbehbahani F, Hoghooghi Esfahani H and Bajoghli S. A comprehensive multidimensional analysis of mental health challenges in the digital age. *Int J Web Res* 2025; 8: 25–48.